DOCUMENT RESUME

ED 047 559 FL 001 984

AUTHOR	Schrand, Heinrich, Ed.
TITLE	Leistungsmessung im Sprachunterricht: Positionspapier (Measuring Achievement in Language Instruction: A Position Paper).
INSTITUTION	Informationszentrum fur Fremdsprachenforschung, Marburg (Germany).
SPONS AGENCY	Deutscher Volkshochschul-Verband, Frankfurt (West Germany).
PUB DATE	69
NOTE	151p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$6.58
DESCRIPTORS	Achievement Tests, Aptitude Tests, Certification, Educational Certificates, Evaluation, Evaluation Methods, *German, Instructional Program Divisions, *Language Instruction, *Language Proficiency, *Language Tests, Linguistic Performance, Performance Criteria, Performance Specifications, *Second Language Learning, Supervisory Methods

ABSTRACT

This position paper on the measurement of achievement in language instruction contains nine articles by leading authorities. Three articles, written in German, cover: (1) testing: necessities and dangers, (2) supervision and evaluation of oral achievement in certification examinations, and (3) development of a higher level of an adult education language certificate for commerce and industry. The articles in English discuss language testing, measurement of oral communication, criterion-referenced testing, oral English testing, error analysis, and picture composition and oral-response tests. (RL)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED0 47559

IFS Dokumentation

Leistungsmessung im Sprachunterricht

Positionspapier

Herausgegeben in Zusammenarbeit mit der Pädagogischen Arbeitsstelle des Deutschen Volkshochschul-Verbandes vom Informationszentrum für Fremdsprachenforschung
355 Marburg/Lahn, Liebigstraße 37, Telefon 06421/692141

FL 001 984

IFS

LEISTUNGSMESSUNG IM SPRACHUNTERRICHT

ZWEITE INTERNATIONALE EXPERTENKONFERENZ ÜBER TESTMETHODEN IM FREMDSPRACHENUNTERRICHT DER ERWACHSENENBILDUNG

ZUSAMMENGESTELLT UND BEARBEITET VON

HEINRICH SCHRAND

INFORMATIONSZENTRUM
FÜR FREMDSPRACHENFORSCHUNG
MARBURG/LAHN 1969

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Informationszentrum für Fremdsprachenforschung TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

Copyright by INFORMATIONSZENTRUM FÜR FREMDSPRACHENFORSCHUNG

INHALT

VORWORT 5

TESTEN IM SPRACHUNTERRICHT - NOTWENDIGKEITEN UND GEFAHREN
Reinhold Freudenstein, Informationszentrum für Fremdsprachenforschung, Marburg 8

LANGUAGE TESTING
Alan Davies, University of Edinburgh 27

MEASUREMENT OF ORAL COMMUNICATION
John A. Upshur, University of Michigan, Ann Arbor 53

CRITERION-REFERENCED TESTING
Rebecca M. Valette, Boston College, Boston 81

ORAL ENGLISH TESTING
Roy H. Kingsbury, European Language and Educational Centres, Bournemouth 103

ERROR ANALYSIS
Alan Davies, University of Edinburgh 109

ZUR KONTROLLE UND BEWERTUNG MÜNDLICHER LEISTUNGEN IN DER ZERTIFIKATSPRÜFUNG
Heinrich Schrand, Informationszentrum für Fremdsprachenforschung, Marburg 125

PICTURE COMPOSITION AND ORAL-RESPONSE TESTS
Terence Creed, Royal Society of Arts, London 131

ZUR ENTWICKLUNG EINER HÖHEREN STUFE DES VOLKSHOCHSCHUL-SPRACHZERTIFIKATS FUR DEN BEREICH WIRTSCHAFT UND INDUSTRIE
Robert Nowacek, Deutscher Volkshochschul-Verband, Frankfurt 149

VORWORT

Die Reformbestrebungen des Deutschen Volkshochschul-Verbands auf dem Fremdsprachensektor haben seit Jahren intensive Impulse durch die der Leitung von EBERHARD PIEPHO unterstellten Internationalen Neusprachler-Tagungen erhalten. Angesichts dieser engen Bindungen schien es naheliegend, die vom Deutschen Volkshochschul-Verband in Zusammenarbeit mit dem Europarat veranstaltete Zweite Internationale Expertenkonferenz über Testmethoden im Fremdsprachenunterricht der Erwachsenenbildung in die vom 23. April bis 2. Mai 1969 stattfindende XII. Internationale Neusprachler-Tagung im Haus Sonnenberg bei St. Andreasberg zu integrieren.

Die Diskussion und der Gedankenaustausch, die durch die Hauptreferate ausgelöst wurden, erreichten damit einen elitären Kreis, der repräsentativ für alle Bereiche des Fremdsprachenunterrichts von der Volksschule bis zur Universität war. Gemeinsamer Tenor der Diskussionen war die Erkenntnis, die fortschrittlichen und hervorragend bewährten Testmethoden im Fremdsprachenunterricht der deutschen Erwachsenenbildung sollten auch im Bereich des öffentlichen Schulwesens Eingang finden. Ebenso deutlich wurde die Forderung, an den Universitäten und Pädagogischen Hochschulen in der Bundesrepublik der Grundlagenforschung auf dem Gebiet des *language testing* wesentlich stärkere Beachtung zu schenken und ein zentrales Testinstitut zu gründen.

Die auf dem Sonnenberg ausgelöste Diskussion zog eine lebhafte Reaktion nach sich: Aus allen Teilen der Bundesrepublik und aus den verschiedensten Schulbereichen gingen bei der Bundesprüfungszentrale des Deutschen Volkshochschul-Verbands Anforderungen auf Kopien einzelner Referate ein.

Das legte es nahe, einen umfassenden Bericht über die Konferenz in gedruckter Form vorzulegen. Angesichts der gemachten Erfahrungen sollte die Publikation jedoch nicht nur den Mitarbeitern in der Erwachsenenbildung zugänglich sein. Da durch jahrelange enge Zusammenarbeit mit REINHOLD FREUDENSTEIN der Deutsche Volkshochschul-Verband in ihm einen der konsequentesten Verfechter einer Reform des Fremdsprachenunterrichts im deutschen Schulwesen kennengelernt hatte, trat er an ihn mit der Bitte heran, diesen Bericht in die Reihe der Publikationen des Informationszentrums für Fremdsprachenforschung zu übernehmen. Ihm und seinen Mitarbeitern - vor allem HEINRICH SCHRAND, der für die Redigierung verantwortlich zeichnete - sei an dieser Stelle herzlich gedankt.

Möge dieser Bericht dazu beitragen, den Forderungen der Sonnenberg-Konferenz auf allen Ebenen des Fremdsprachenunterrichts zum Durchbruch zu verhelfen.

August 1969

Robert Nowacek

Bundesprüfungsbeauftragter für
Fremdsprachen im
Deutschen Volkshochschul-Verband

TESTEN IM SPRACHUNTERRICHT - NOTWENDIGKEIT UND GEFAHREN

Reinhold Freudenstein

Wenn sich Experten zu einer Konferenz mit dem thematischen Schwerpunkt "Testen" versammeln, dann liegt leicht die Gefahr nahe, daß sie über der Erörterung spezieller Einzelfragen und bestimmter Probleme die Wirklichkeit der Leistungsmessung in der täglichen Praxis aus dem Blick verlieren. Es ist darum notwendig, zunächst übersichtlich darzustellen, wie es um Testen, Tests und Testverfahren im Fremdsprachenunterricht der deutschen Schulwirklichkeit bestellt ist. In einem zweiten Teil sollen dann Fragen erörtert werden, die die Leistungen des Tests im Fremdsprachenunterricht kritisch analysieren; die Anerkennung des Tests als Prüfungsform ist nicht unproblematisch, und es müssen bestimmte unterrichtliche Voraussetzungen vorliegen, um den didaktischen Ort des Tests festlegen zu können, damit der objektiven Leistungsmessung ein Entfaltungsraum geboten werden kann.

Dieser kritische Ansatz macht es notwendig, ein "Bekenntnis" an den Anfang zu setzen: Ich bin für den Test. Ich glaube, daß nur durch den Test eine objektive Überprüfung der Schülerleistungen möglich ist, und ich bin fest davon überzeugt, daß der Test einen Beitrag zur Demokratisierung der Schule und zum Abbau subjektiver Strukturen in der Leistungsbeurteilung durch den Lehrer zu leisten vermag. Diese Feststellung ist notwendig, weil in der detaillierten Analyse leicht der Eindruck entstehen könnte, daß die Kritik überwiegt und der Test darum überflüssig sei. Wenn der Test hier kritisch betrachtet wird, dann nur aus dem einen Grunde, um von Anfang an zu versuchen, ihn in Form und Inhalt optimal zu gestalten. Diese "Kritik" ist darum in Wirklichkeit ein Beitrag zu seiner Vervollkommnung und nicht ein Argument, das gegen ihn spricht. Wir üben also Kritik am Test, weil er uns am Herzen liegt.

1. ZUR TESTSITUATION IN DER SCHULWIRKLICHKEIT.

Man kann die Situation und das Problem der Anerkennung von Tests und Testverfahren im Sprachunterricht der Bundesrepublik nicht besser kennzeichnen als durch einen Hinweis auf die Fachliteratur. Heft 4 der Schriftenreihe *Der Fremdsprachliche Unterricht*[1] stand unter dem Thema *Schriftliche Arbeiten im Neusprachlichen Unterricht*. Auf 80 Druckseiten wurde das Wort Test nicht einmal erwähnt. Auch im Sachregister der bekannten und weitverbreiteten Methodik von SCHUBEL[2] sucht man in der vierten, neu bearbeiteten Auflage von 1964 das Stichwort *Test* vergeblich. Die fremdsprachliche Methodik stützt sich nach wie vor auf sogenannte "bewährte" Prüfungsformen: das Diktat, die Nacherzählung, die grammatische Arbeit, die Vokabelarbeit, das précis, die Übersetzung - und das nicht nur trotz LADOS *Language Testing*[3], sondern geradezu in Auflehnung gegen ihn. Der einzige, der sich intensiver mit dem Test im neusprachlichen Unterricht befaßt hat, war LEISINGER (1966), und in seinem Buch *Elemente des Neusprachlichen Unterrichts*[4] kommt er zu dem Schluß, die Testvorschläge LADOS seien kaum geeignet, ein harmonisches, ausgeglichenes Bild der Schülerleistung zu ermitteln. Er wendet sich scharf gegen die "Addition noch so verläßlicher Einzelprüfungen spezieller Elemente", d.h. also gegen den Test überhaupt, und er verteidigt die üblichen Arbeitsformen - jene global orientierten Komplexprüfungsformen - die bereits um 1900 in der Schule verwendet wurden und die sich auch heute noch zäh behaupten.

Für die Mehrzahl unserer Fremdsprachenlehrer ist der Test bis heute eine mehr oder weniger unseriöse Angelegenheit. Zu dieser Einstellung haben zwei Faktoren entscheidend beigetragen. Einerseits begegnet man dem Wort *Test* ständig im Raum der Massenmedien; man kann in Illustrierten testen, ob man eine gute Hausfrau ist, ob man seine Kinder richtig erzieht oder zum Partner paßt. Das geschieht - ebenso wie auch bei der Führerscheinprüfung - durch das *Auswahl-Antwort-Verfahren*, also den *multiple-choice test*. Das Wort *Test* und der Vorgang des Ankreuzens von einem unter drei, vier oder fünf Kästchen sind daher in der Vorstellung vieler Fremdsprachenlehrer synonym. Es fehlt - und das ist der zweite Faktor - an der notwendigen Sachkenntnis. In der Universitätsausbildung wird sie in der Regel nicht vermittelt, und wer sich in der Literatur orientieren möchte, stößt allenthalben auf Verständnisschwierigkeiten. Hat man schließlich die Terminologie von *Korrelationskoeffizient* über *Reliabilität*, *Vertrauensgrenzen* und *Median* bis hin zum *Profil* erarbeitet, stößt man auf die Notwendigkeit mathematischer und statistischer Kenntnisse, und hier scheitern meist die letzten, die guten Willens sind.

Es ist also richtig, wenn KARLHEINZ INGENKAMP sagt, Deutschland sei auf dem Gebiet des *educational measurement* ein *unterentwickeltes Land*.[5] Und wenn es dann weiter heißt, ein differenziertes Schulwesen in einer demokratischen Gesellschaft könne auf den Test nicht verzichten, dann liegen in einer solchen Äußerung die Wurzeln für eine weitere Begründung, warum der Test so vielen Lehrern suspekt ist. Hier wird das neue Prüfungsverfahren nämlich gleichgesetzt mit anderen Bemühungen einer Demokratisierung der Schule im Zuge von Reformen und Neuordnungen, die gerade in der Lehrerschaft unseres reformfreudigen Föderalismus nicht auf allzu große Gegenliebe stoßen. Es wird darauf verwiesen, daß die Fülle der Neuerungen und Verordnungen für den Schulalltag kaum noch überschaubar sei, und viele Lehrer beschränken sich aus diesem Grunde darauf, zunächst einmal abzuwarten, bis sich ein neues Verfahren fest etabliert hat - eine Auffassung, die von der Notwendigkeit zur permanenten Reform des Schulwesens, und darin auch der Fachdidaktik, weit entfernt ist. Am aufgeschlossensten sind für Fragen des Tests lediglich diejenigen Lehrer, die moderne Medien in den Fremdsprachenunterricht einbeziehen. PIMSLEUR hat gesagt, je ausgeprägter die audiolinguale Methode in einer Klasse praktiziert werde, desto notwendiger wäre der Einsatz spezialisierter Tests.[6] Die audiolinguale Konzeption herrscht heute dort vor, wo man der Sprache als Kommunikationsmittel mehr Bedeutung zuerkennt als ihrem Beitrag zur sogenannten "formalen Bildung"; und das geschieht meist in Verbindung mit technologischen Medien, die die Sprachfertigkeiten ausbilden helfen. Darum sind es in erster Linie die Sprachlaborlehrer, die sich gegenwärtig am intensivsten mit der Frage des Testens in der täglichen Praxis befassen. Wegbereiter für eine systematische und objektive Leistungsmessung im Sprachlabor sind KING, MATHIEU, HOLTON, STACK und LADO, und es liegen auch bereits einige deutsche Arbeiten vor, die sich mit speziellen Untersuchungen zum *monitoring* als Test befassen. Aber auch auf diesem Gebiet gilt, daß das Sprachlabor längst nicht in dem Maße, wie es technisch möglich, pädagogisch sinnvoll und fachmethodisch vertretbar wäre, als Testraum entdeckt worden ist.

Auch der Ausblick auf die Zukunft ist nicht allzu ermutigend. Auf dem Gebiet von Forschung und Lehre müssen zunächst die Versäumnisse der Vergangenheit nachgeholt werden, ehe daran gedacht werden kann, auf breiter Basis eigene Entwicklungen voranzutreiben. Die in den Bundesländern gültigen Prüfungsbestimmungen fordern den Test keineswegs als verbindliches Element im Prüfungsprozeß, und er gehört noch immer - aus ministerieller Sicht - zu den schulischen Experimentalformen. Im Unterricht der Schule werden also auch in den kommenden Jahren lediglich diejenigen Schüler vom Test und seinen Vorzügen profitieren,können, die von jener relativ kleinen Gruppe interessierter und überzeugter Lehrer unterrichtet werden, denen ihr eigenes Urteil suspekt geworden ist und die sich nach neuen Möglichkeiten zur Objektivierung der Schülerleistungen umsehen.

Es gibt - und das soll nicht verschwiegen werden - nicht nur diesen negativen Aspekt. Ansätze, sich den Fragen des Testens im Sprachunterricht zu stellen, finden wir an mehreren Stellen. Die Prüfungsbestimmungen zum Zertifikat *Englisch* und *Französisch* des *Deutschen Volkshochschul-Verbandes* setzen den Test als Prüfungsform bereits voraus; diese Zertifikatsbestimmungen sind darum ein wichtiger Meilenstein in der Entwicklung von Tests für den Fremdsprachenunterricht in Deutschland. Auch einige Verlage beginnen zögernd mit der Herausgabe informeller Leistungstests zur Überprüfung erworbener Kenntnisse nach der Durcharbeitung eines Kapitels oder eines ganzen Lehrbuchs. Auch im Zusammenhang mit Schulversuchen werden Tests im Fremdsprachenunterricht eingesetzt.

Die integrierten Gesamtschulen des Landes Hessen haben im Juli 1969 einen lehrbuchunabhängigen Leistungstest eingesetzt, der die Grundlage für die Einteilung der Schüler in Niveau- und Kursgruppen bildete. In der Reihe *Deutsche Schultests* sind einige Tests in Anlehnung an Englischlehrbücher erschienen[7], und schließlich beginnen auch an einigen Universitäten und Pädagogischen Hochschulen Lehrveranstaltungen, die sich mit der objektiven Leistungsmessung beschäftigen.

Möglicherweise wird es sich einmal als vorteilhaft erweisen, daß ein relativ langsamer Anlauf der Testarbeit im Fremdsprachenunterricht zu beobachten ist; auf diese Weise können sich die neuen Verfahren vielleicht intensiver in das Bewußtsein derer einpflanzen, die den Test eines Tages zu administrieren haben. Wird das Neue zunächst kritisch reflektiert, ehe es praktisch appliziert wird, kommt dies der Sache, um die es geht, zugute. Denn auch dies muß gesagt werden: Wir sind noch längst nicht so weit, daß wir mit Sicherheit und Überzeugung behaupten können, daß der Test pauschal an die Stelle bisheriger Prüfungsformen treten und das gleiche bzw. mehr als traditionell gegliederte Verfahren leisten könne. Ganz bewußt vertreten darum auch die Testspezialisten in der Bundesrepublik die Ansicht, der Test bewähre sich in Ergänzung, also als additives Element zu anderen Prüfungsformen. Auch INGENKAMP behauptet, es wäre pädagogisch verantwortungslos, auf den Test als Hilfsmittel, und d.h. zur Ergänzung des eigenen, subjektiven Urteils zu verzichten. Es wäre also ein Mißverständnis, wollte man davon ausgehen, daß der Test der einzige Maßstab für die Beurteilung einer Schülerleistung sein könnte.

Aber auch unter diesen einschränkenden Gesichtspunkten gilt, daß die pädagogische Öffentlichkeit noch nicht davon überzeugt ist, daß Tests die gängigen Beurteilungsverfahren wesentlich ergänzen können. Wer für den Test eintritt, muß darum nachweisen können, nicht nur daß, sondern genau an welchen Stellen der Test unproblematischer, genauer, weniger subjektiv, vielleicht sogar wirklich objektiv ist, kurz, eine bessere Prüfungsform darstellt als das, was wir bisher praktizieren. Und an dieser Stelle beginnen sich sogleich und sehr massiv ernste Probleme einzustellen.

2. PROBLEME DER OBJEKTIVEN LEISTUNGSMESSUNG.

Ohne damit den Anspruch auf Vollständigkeit erheben zu wollen, sind es drei Aspekte, von denen aus der Test kritisch in Frage gestellt werden kann. Diese Aspekte sind von grundsätzlicher Bedeutung, und ihr Schatten spiegelt sich in jeder einzelnen Testaufgabe, die wir zu schreiben oder zu administrieren haben.

Sie werden hier, und das sei nochmals betont, nicht erörtert, um gegen den Test als solchen zu polemisieren; sie sollen lediglich dazu anregen, das Bewußtsein um die Testproblematik kritischer zu reflektieren, als die weithin fraglose Anerkennung bisheriger Prüfungsformen, die ja nicht deshalb so weit verbreitet sind, weil sie den pädagogischen oder didaktisch-methodischen Anforderungen genügen, sondern weil sie ganz einfach als gegeben hingenommen werden.

a. *Der Verzicht auf das individuelle Prüfungsmoment.*

In der Literatur wird immer wieder darauf verwiesen, daß es einer der größten Vorzüge des Tests sei, individuelle Momente, subjektive Faktoren, also: die aufs Ganze gesehen sehr zufälligen Beurteilungskriterien einzelner Lehrer überwinden zu helfen. Es ist bekannt, daß weder die langjährige Erfahrung eines Lehrers noch ein klasseninterner Maßstab geeignet sind, dem Schüler das Vertrauen zu geben, er werde richtig und gerecht beurteilt.

CHAUNCEY berichtet über einen zu dieser Frage angestellten und kontrollierten Versuch, der deutlich erkennen läßt, daß die Lehrer bei der Beurteilung weder mit sich selbst übereinstimmen noch sich untereinander einig sind.[8]

So wurde bei einer Untersuchung ein Aufsatz aus der 8. Klasse von 28 Lehrern zweimal gelesen. Fünfzehn von ihnen, die ihn beim ersten Durchgang mit *ausreichend* oder besser bewertet hatten, hielten ihn beim zweiten Lesen für nicht ausreichend; elf Lehrer, die ihn zunächst als *mangelhaft* bezeichnet hatten, bewerteten ihn später positiv. In einem anderen Fall erhielt ein Aufsatz, der von 142 Lehrern gelesen wurde, zwischen 50 und 98 Punkte.

Es ist also nicht möglich, Bewertungsdifferenzen zu verhindern; und von dieser Tatsache leben nicht zuletzt auch die Spalten der methodischen Fachliteratur, in denen unter der Überschrift *Would you have marked it wrong*, oder *Ist das gutes Deutsch?* Schülerfehler als richtig und die Lehrerkorrektur als falsch entlarvt werden.

Diese Tatsache der subjektiven Beeinflussung der Bewertung - und sie beginnt schon bei der Auswahl des Prüfungsstoffes und bei der Entscheidung für eine bestimmte herkömmliche Prüfungsform - ist unbestritten. Andererseits ist zu fragen, inwieweit Objektivierung und Standardisierung, so wesentlich sie sind, nicht zu einer Nivellierung führen, und zwar zu einer Nivellierung der Leistungsanforderungen, da schließlich nur noch das gelehrt werden könnte, was objektiv überprüfbar ist. Nimmt man das individuelle Moment aus der Prüfung, so nimmt man es im Grunde auch aus dem Unterricht. Ein Lehrer, der unter dem Zwang enggefaßter, von außen gegebener Kriterien zu unterrichten hat, kann sich nicht mehr in der Weise engagieren, die seinen persönlichen Interessen, seinem persönlichen Stil und seinen persönlichen Gaben entspricht. Der Zwang vorgegebener Lehrzielkriterien, bis ins einzelne präzisiert und im Test dann im einzelnen abgerufen, würde zurückdrängen können, wofür er eigentlich als Modell dienen könnte, wenn man MORTON folgt: als Modell für *imagination, sensitivity, understanding, sympathy, feeling and humility. These are human values for which language almost alone can serve for communication.*[9]

Nicht verschwiegen werden darf in diesem Zusammenhang, daß diese Gefahr auch schon im herkömmlichen Prüfungssystem besteht. Im Bundesland Bayern, das der Übersetzung eine wesentliche Bedeutung im Spracherlernungsprozeß beimißt, obwohl die Übersetzung zur Spracherlernung absolut nichts beiträgt, würde ein Lehrer unverantwortlich handeln, wenn er seine Schüler nicht auf diese offizielle Prüfungsform vorbereitete. Auch dort, wo im Abitur neue Prüfungsformen experimentell erprobt werden, bereiten die Lehrer mit gutem Recht ihre Sprachschüler über Jahre hindurch auf diese Prüfungsformen vor. Der Eindruck einer direkten Abhängigkeit von Prüfungsform und Lehrverfahren verstärkt sich jedoch beim Test insofern, als hier nicht mehr globale Bildungsziele pauschal überprüft werden, sondern spezielle Einzelinformationen einer detaillierten Analyse unterzogen werden.

Und noch ein weiteres sei erwähnt: daß die schulische Note im Fremdsprachenunterricht nicht die Addition von Einzelleistungen allein widerspiegelt. Reduzierte man das Zeugnis unserer Schüler zu einem reinen Leistungsreport, gäbe man damit auch den Anspruch auf einen pädagogischen Auftrag der Schule auf. Sicherlich werden sich immer Beispiele für Willkür in der Notengebung finden lassen; es lassen sich daneben auch ebenso viele Beispiele für eine bewußte

Steuerung schulischer Notengebung finden, die gut begründet werden können und die sehr oft zu dem gewünschten Ziel geführt haben.

Wesentlich ist: das Testergebnis ist seinem Wesen nach unpersönlich; es ist korrekt, objektiv und richtig, aber zweierlei kann es nicht leisten: 1. Es erfaßt den Schüler nicht als Gesamtpersönlichkeit, es berücksichtigt nicht individuelle Faktoren, von denen eine Leistung immer beeinflußt wird und die der verantwortungsbewußte Lehrer immer in Rechnung setzt; 2. Es versagt weithin bei der *Beurteilung integrierter Fertigkeiten, immer bei der Beurteilung von komplexen Fähigkeiten.* Diese Aussage ist wesentlich für den zweiten Gedankenkreis.

b. *Was lehren wir, was testen wir?*

Der Test wurde in der Bundesrepublik im Zusammenhang mit einer Reform des neusprachlichen Unterrichts aktuell, die vornehmlich in Verbindung mit technischen Unterrichtsmitteln bisherige Lehrziele und Lehrverfahren in Frage stellt. Diese Reformbewegung konzentriert sich bis heute auf die sogenannten audiolingualen oder audiovisuellen Konzeptionen, bei denen die Sprache als Kommunikationsmittel zunächst ausschließlich im Mittelpunkt des Lehr- und Lernprozesses steht. Wir stehen heute am Anfang dieser Bewegung. Tatsächlich aber wird in den Schulen noch immer eine *vermittelnde* Methode praktiziert - ein Mittelweg zwischen der grammatisierenden Übersetzungsmethode und der direkten Methode, die beide in der Theorie als Extremwege überwunden sind. Es gehört zum Wesen einer vermittelnden Methode, ein breites Lehrzielfeld anzustreben, in dem neben der gesprochenen Sprache auch die geschriebene Sprache, die Lan-

deskunde, die Literatur, formale Kriterien und vor allem die Grammatik einen festen Platz einnehmen. Jeder Lehrer weiß aus seinem eigenen Unterricht, welche Bedeutung der Grammatik und grammatischen Einzelkenntnissen bis hin zu Regeln in unserem Unterricht zukommt. Auch dann, wenn Regeln aus Beispiel- und Mustersätzen deduziert werden - es bleiben Regeln; auch dann, wenn das Lehrbuch noch so schön illustriert und aufgelockert gestaltet ist, sein zugrundeliegendes Gerüst ist immer der künstliche Rahmen einer weithin an der lateinischen Grammatik orientierten Gesetzmäßigkeit. Wir lehren also weithin immer noch Sprache über Grammatik, aber wir überprüfen in der Schulwirklichkeit von heute die Kenntnisse unserer Schüler so, als hätten wir Sprache als Kommunikationsmittel, als Zweitsprache gelehrt, nämlich ganzheitlich. Das Diktat, die Nacherzählung, der Aufsatz - das sind genau die gleichen Prüfungsformen, die auch dem muttersprachlichen Unterricht eigen sind. Wären wir konsequent, müßten wir *das* in den Mittelpunkt der Klassenarbeiten stellen, was auch im Mittelpunkt des Unterrichts steht, und das wäre bis heute die Grammatik.

Natürlich kann man einwenden, daß die Grammatik im modernen Unterricht nur Hilfe, nur Stütze, nur Gerüst sei; sie trägt das, worauf es eigentlich ankommt: das Sprechen, Lesen, Schreiben und Analysieren in einer fremden Sprache. In der Theorie mag das stimmen; in der Praxis jedoch bleibt bis hin zum Abitur das grammatische Gerüst im Vordergrund jedes Bemühens, und konsequenterweise zählen die grammatischen Fehler am schwersten, und zwar auch dann, wenn es eigentlich auf sie gar nicht ankommen dürfte, wie z.B. in der Nacherzählung. Die Argumentation dieses Gedankenkreises läßt sich vereinfachend so darstellen:

Im bisherigen Fremdsprachenunterricht haben wir spezielle Einzelkenntnisse gelehrt und mit überkommenen Prüfungsformen die Sprache als Kommunikationsmittel getestet, in der methodischen Neuorientierung lehren wir die Sprache als Kommunikationsmittel, testen aber spezielle, detaillierte und in sehr vielen Fällen für den Kommunikationsvorgang überhaupt nicht relevante Einzelerscheinungen.

Noch einfacher ließe sich behaupten: Wir haben Grammatik gelehrt und Sprache getestet - jetzt lehren wir die Sprache und testen Grammatik. Der Test müßte also in erster Linie *gesprochene* Sprache überprüfen; die bisher vorliegenden Fremdsprachentests beschränken sich jedoch auf eine Überprüfung der Schreib- und Lesefertigkeiten, und das gesprochene Wort kann nur durch sehr ungenügende indirekte, objektive Aushilfsverfahren getestet werden. LADO gesteht in *Language Testing: The testing of speaking is less advanced than the testing of listening, and although we talk a great deal about the pronunciation of students of foreign language, we do relatively little about testing it.*[10]

Das liegt natürlich nicht an einer grobfahrlässigen Vernachlässigung seitens der Testkonstrukteure, sondern ganz einfach daran, daß die gesprochene Sprache nicht ebenso einfach, sicher und objektiv überprüfbar ist wie das Hören, das Schreiben und das Lesen. Hier stoßen wir an eine wesentliche Grenze des Tests im neusprachlichen Unterricht. Mancherorts ist man darangegangen, diese Grenze zu überwinden, und zwar durch die Entwicklung von Maschinen, die eine automatisch gesteuerte Antwortkontrolle ermöglichen. Bisher zeigt sich jedoch, daß ein solcher Schritt zur Objektivierung mündlicher Schülerleistungen einen Aufwand erfordert, der in keinem Verhältnis zum erwarteten Erfolg steht.

Und weiter: Machten wir wirklich ernst mit dem Bemühen, Sprache als Kommunikationsmittel zu lehren und zu testen, dann müßten sich die Kriterien zur Fehlerbewertung an völlig anderen Kategorien orientieren, als sie auch heute noch den Tests zugrundeliegen. Wir dürften nämlich nicht mehr fragen, ob Massachusetts oder Mississippi mit der richtigen Anzahl von s-Buchstaben geschrieben ist, ob das richtige Pronomen gewählt wurde, ob die Zeitenfolge stimmt usw., sondern die Leitfragen müßten lauten: Hat der Schüler den fremdsprachlichen Text verstanden? Reagiert er spontan und inhaltlich richtig, ist das, was er sagt, verständlich? Oder, mit einer Formulierung von GERHARD NICKEL, *does the message get across?* Natürlich macht eine Dänin, die sich um eine Sekretärinnenstelle beworben hat und am Telefon fragt: *Wann kann ich Ihnen sprechen?* einen grammatischen Fehler, und überdies kann auch die Intonation noch falsch sein, aber diese Fehler beeinträchtigen die Kommunikation keineswegs, und sie sind darum höchstens Schönheitsfehler, keineswegs aber grammatische Vergehen, vor denen man Angst haben müßte. Die mit diesem Beispiel angeschnittenen Fragen und Probleme müssen im einzelnen angegangen und gelöst werden, wenn der Test zum integrierenden Bestandteil eines fremdsprachlichen Lehrgangs gehören soll.

c. *Von der Prüfungsangst zur Testpsychose.*

Es gehört zu den wichtigen Anliegen der Erziehungswissenschaft in vielen Fächern, dafür zu sorgen, daß die Angst- und Stress-Situation unserer Schüler vor einer Klassenarbeit, also die Prüfungsangst, überwunden werden kann. Patentlösungen gibt es bis heute nicht, und weder ein sozialintegratives Eingehen des Lehrers auf die Fragen und

Wünsche seiner Schüler während der Klassenarbeit, noch das Überrschungsmoment der vorher nicht angekündigten Arbeit haben geholfen, die teilweise entsetzliche Angst der Schüler vor den Arbeiten abzubauen.

Diese Behauptung wird durch eine Untersuchung von HORST HAECKER bestätigt, der empirisch belegt hat, daß die Schülerangst vor den Klassenarbeiten die Entwicklung der Persönlichkeit störe und Konflikte schaffe, während es doch eigentlich nur darum geht, den Stand von Leistungen aufzuweisen. Mit mehr als 800 Schülern und Schülerinnen im Alter von 9 - 14 Jahren hat HAECKER Versuche angestellt. Dabei ergab sich, daß der Leistungsdruck, der mit einer Klassenarbeit verbunden ist, das Arbeitsresultat nachteilig beeinflußt. Im Anfang der Schulzeit, so stellte sich heraus, sind die Kinder noch unbefangen; in den ersten beiden Grundschulklassen waren die Leistungen, die sie bei einer Prüfungsarbeit zeigten, sogar besser als die bei einer bloßen Übungsarbeit. Vom 3. Schuljahr an aber zeigte sich bei dem Test- oder Prüfungsdruck eine deutliche Tendenz zur Leistungsminderung. Mehr als die Hälfte der Schulkinder gab an, eine Arbeit "gern" zu schreiben, zugleich aber klagten sie darüber, daß sie so "aufgeregt" seien. Mädchen litten offenbar mehr unter der Prüfungsangst als Jungen. Zudem ergab sich, daß die Aufregung beim Rechnen am größten, beim Aufsatzschreiben am geringsten ist. Aus seinen Untersuchungen zog HAECKER den Schluß, die Pädagogen sollten bei der Beurteilung der Arbeiten die schwierige Konfliktsituation der Schüler, ihr Schwanken zwischen Freude an der Sache und an der Herausforderung und andererseits der Furcht vor Fehlern, Versagen und Strafe in Rechnung stellen. Die übliche einseitige Bewertung der Zensuren für schriftliche Arbeiten, so meint der Psychologe, könne sach- und zielfremde Motive beim Lernen begünstigen: Notenehrgeiz und Neid, Angst und Minderwertigkeitsgefühle.

Merkwürdigerweise gelten diese Aussagen nicht nur für die herkömmlichen Arbeiten, sondern auch für den Test. Trotz aller Bemühungen, die wir im Zusammenhang mit der neuen Prüfungsform anstellen, um diese Angst zu bannen, tritt sie vor einem Test wesentlich größer, erheblich verstärkter auf als vor herkömmlichen Prüfungen. Auch kleine *Tricks* helfen kaum, der Testpsychose Schranken zu setzen - indem wir z.B. mehrfach anmerken, es handele sich um eine Übung, wir wollten einmal feststellen, wie gut wir unterrichtet hätten, ob das Unterrichtsmaterial geeignet sei usw.

Woran liegt das? In erster Linie wohl daran, daß die Schüler den Zwang zur Objektivität, die Unmöglichkeit, sich diesem Zwang in irgendeiner Weise zu entziehen, spüren. Die Schüler erfahren - bewußt oder unbewußt - die Einsicht in die Unentrinnbarkeit der individuellen Unvollkommenheit vor der standardisierten Perfektionierung. Hier wird ganz klar beurteilt, indem gezählt, gemessen, korreliert wird; hier ist niemand mehr gleich gut oder gleich schlecht wie ein anderer, sondern jeder erhält einen Rangplatz, den er jederzeit verlieren kann, den er natürlich auch verbessern kann, aber nur, wenn er das, was als nächstes getestet wird, mehr oder weniger zufällig weiß oder nicht. Diese Formulierung ist bewußt scharf und übertrieben, weil wir uns selbst immer wieder sagen müssen, daß wir noch längst nicht wissen, ob das, was wir für einen Test inhaltlich auswählen, wenn wir unsere *items* schreiben, tatsächlich auch diejenigen Inhalte sind, die es andererseits erlauben von einer *objektiven* Überprüfung der Schülerleistungen zu sprechen. Weithin ist bis heute auch die Auswahl dessen, was wir testen, zufällig und subjektiv - und die Leistungen, die wir auf dieser Grundlage ermitteln, sind es ebenso, auch wenn sie in sich stimmen.

Es ist durchaus wichtig und notwendig, den Test als integrierten Bestandteil in unser Lehrverfahren aufzunehmen; es wäre falsch, ihn an die Stelle herkömmlicher Prüfungsmethoden zu setzen, und es wäre gefährlich, zu einem so frühen Zeitpunkt der Entwicklung dieses Arbeitsfeldes seine Bedeutung zu überschätzen. Daraus ergibt sich die Folgerung, daß der Test zunächst nur als Ergänzung herangezogen werden darf, wobei wir uns selbst immer wieder sagen sollten, daß wir den Stellenwert dessen, *was* wir mit dem Test prüfen und *wie* wir es prüfen, noch nicht kennen - noch nicht so genau jedenfalls, daß wir die Testergebnisse unkritisch und anbetungsvoll als gegebene Objektivität hinnehmen könnten.

Der Einsatz objektiver Testverfahren im Fremdsprachenunterricht der allgemeinbildenden Schule und auch der Erwachsenenbildung wird nicht nur deshalb notwendig werden, weil mit steigenden Schülerzahlen in Zukunft eine individuelle Bewertung schriftlicher Arbeiten kaum noch möglich sein wird; nur durch den Test ist es rationell, individuell und effektiv gewährleistet, den Leistungsstand der Schüler zu messen und damit Grundlagen zu erstellen, die eine wesentliche Voraussetzung für das Nachholen von Versäumtem und die weitere Arbeit in einer ganz bestimmten, jeweils individuell verschiedenen Richtung sind. Die Aufgaben, die im Testbereich auf uns zukommen, bestehen also darin, die Mängel bisheriger Prüfungsformen zu beseitigen und gleichzeitig dafür zu sorgen, daß mit neuen Formen der Leistungsmessung nicht neue Probleme auftreten, deren Bewältigung schwieriger ist als eine mögliche Korrektur traditioneller Meßverfahren. Eines scheint schon heute sicher zu sein, und es sei als These an den Schluß dieser Ausführungen gestellt. Wer sich vorurteilslos mit Fragen der Leistungsmessung im Fremdsprachenunterricht befaßt, der wird feststellen, daß ein heute noch unvollkommener Test im Blick auf das, was er leisten soll und zu leisten vorgibt, in jedem Falle mehr leistet und darum besser ist als eine Prüfungsform herkömmlicher Art. Ausgehend von dieser Hypothese dürfte es nicht schwerfallen, Formen und Typen objektiver Leistungsmessung zu entwickeln, die dazu beitragen, den Lernprozeß der Fremdsprachenschüler optimal zu gestalten.

Anmerkungen:

[1] *Der fremdsprachliche Unterricht (Wissenschaftliche Grundlegung - Methodische Gestaltung).* Herausgegeben von Paul Hartig. Jahrgang 1, Heft 4, Klett Verlag, Stuttgart 1967.

[2] Friedrich Schubel: *Methodik des Englischunterrichts für Höhere Schulen.* Diesterweg Verlag, Frankfurt/M. 1966.

[3] Lados *Language Teaching* liegt in deutscher Übersetzung vor unter dem Titel *Moderner Sprachunterricht* (Hueber Verlag, München 1967); *Language Testing* ist noch nicht übersetzt worden.

[4] Fritz Leisinger: *Elemente des neusprachlichen Unterrichts.* Klett Verlag, Stuttgart 1966.

[5] Karlheinz Ingenkamp: *Die Feststellung der Fremdsprachen-Leistung und -Eignung durch Lehrerurteile und Tests.* In: *Praxis des neusprachlichen Unterrichts,* 14. Jahrgang, Heft 3, Lensing Verlag, Dortmund 1967, S. 277 ff.

[6] Paul Pimsleur: *Ein Sprachen-Eignungstest. Ziele und Durchführung.* In: *Praxis des neusprachlichen Unterrichts,* 14. Jahrgang, Heft 2, Lensing Verlag, Dortmund 1967, S. 134 ff.

[7] Vgl. den Sonderprospekt *Deutsche Schultests,* Verlag Julius Beltz, 694 Weinheim/Bergstraße, Bahnhofstr. 10. Im Verlag Beltz erscheinen auch kritisch-theoretische Arbeiten zu unserem Thema sowie Tagungsberichte (z.B. Bericht über die erste Internationale Arbeitstagung *Möglichkeiten und Grenzen der Testanwendung in der Schule*).

[8] Henry Chauncey: *The Plight of the English Teacher.* In: *The Atlantic Monthly,* Vol. 204, No. 5, p. 123, Nov. 1959.

[9] F. Rand Morton: *The Language Laboratory as a Teaching Machine.* In: *International Journal of American Linguistics* Vol. 26, pp. 113 - 166, 1960.

[10] Robert Lado: *Language Testing. The Construction and Use of Foreign Language Tests.* Longmans, Green, London ^{4}1965, p.78.

LANGUAGE TESTING

Alan Davies

1. AIMS AND PURPOSES

1.1 *Research*

We start with a quotation from a handbook on elementary statistics, a programmed textbook which lives up to its claim to be self-instructional (*McCollough 1963*); the quotation relates to comparisons between groups: *the t test is the most commonly used method of marking this kind of test, and it may be that testing the difference between two means (or averages) is the most common kind of problem met with in psychology and social science* (p. 233). While it may be useful for language teachers to know about tests it is quite unimportant that they should do so. What does matter is that they should recognise that here we are dealing with the first aim of language testing. A test is an index of significance made by performing certain statistical procedures on the results of tests, language or otherwise. The importance of *significance* is critical in research and experiment in education. Language tests are essential for such research to be carried out.

The first aim, then, of language testing is that of *research*. Whether progress or developmental situations are being considered, in the order of the laboratory or in natural settings, tests will at some point be necessary to check on the learning that has gone on so far. Whether it is theories of learning or theories of language, methods of teaching or teaching materials that are being researched into, at some point it will almost certainly be necessary to compare *experimental* and *control* (or *normal*) conditions in order to examine one's research hypothesis. For this purpose it will be necessary to characterise or identify the two approaches, the new and the old, in some way. Tests are useful and necessary ways of doing so. Language tests, then, must be seen first as tools of research in experimental work in education.

1.2 *Progress*

Most language teachers are not likely to be involved in full scale research work. This is true of the teachers within one school and of the teachers in a total educational system. Tests or examinations are just as necessary here but for a different reason from research. The point of a *research* use test is to compare materials, methods etc., with the possibility of changing them; the point of tests with a progress aim is still to compare pupils not with one another but with an already determined standard. (This, as it were, is the ideal aim; in practice pupils often are compared with one another and the already determined standard is a fiction.) However, whatever false there may be in the use of tests to measure progress, the fact is clear and the practice well-known and understood. All teachers want to know how their pupils are getting on, to have some idea of whether what they have taught has been learnt.

It seems sensible to take a fairly common sense and liberal approach to tests for progress and to include all kinds of test methods within them, thus the whole range from the final degree examination, lasting several days or weeks, or the administrative civil service exams right down to the simple question in class, from an essay on Sanskrit phonology or TRIER's field theory of semantics or psycholinguistics down to: what is the meaning of *professeur* or which tense should we use here? These are all *progress* tests, inasmuch as the teacher, examiner or questioner wishes to assess how far learners have gone. (As we shall show later the *achievement* use of language tests is very close to this *progress* purpose.)

1.3 *Guide to Curriculum*

In education all (language) tests and examinations exert an influence on the curriculum and on the teaching. This wash-back effect is often bemoaned by teachers, but once given the fact of examinations then all the more reason for insuring that they are good ones, that their wash-back (which is inevitable) is beneficial and that far from stultifying the curriculum and petrifying the teaching they give it real freedom by describing *exactly* where it is meant to be going. A closely detailed and wide ranging *syllabus* is an essential partner of a good exam which itself needs to be illustrative of tried and defined proficiency levels. There are three opponents to such an approach: one is the traditional one which gives no syllabus save e.g. *there will be a translation into French or A Summary An Oral Interview An Inquiry etc.* This is no help at all: it is as bad as saying: learn French etc. What the teacher wants to know is whose French, how much, what sort and so on. The second opponent is the *Teach the child and the subject* one: this works for skilled and gifted teachers, not for the ordinary teacher whose job is a job and not a vocation. The third opponent is the *Teach through experience of France* *and Frech* one. This is fine (again easiest in the hands of very good teachers) but depends for relevance on a close relationship between the experience that is actually taught in the classroom and the expectations as to what the pupil actually will be able to do when the teaching is over. This leads on to the fourth aim.

1.4 *Representing Terminal Behaviour*

The point of learning a language is presumably to gain control of some kind over that language. Since the *kind* of control projected will presumably be over performance, however limited (and it could be argued that anything less is not worth doing), then the *terminal behaviour* aimed at is that performance, such and such control in such and such situations.

But what does this mean? Some realisation of so vage a notion is needed if syllabus and teaching are to receive the definition they need. The test then, whether along the way or at the final stage of language learning, is a representation in actual terms of these vage notions of terminal behaviour: it stands in their place, and, at the same time acts as an adequate sample of them; the universe of discourse postulated for the learner's control is too vast ever to be completely tested: the test samples the situations, the items and the features which the learner should have grasped. Of course the adequacy of the test itself in its *sampling* is another problem, and a serious one: it is a problem of validity and will be dealt with later.

2. PROGRAMMING INFLUENCE

Programmed instruction, employing its twin approach of analytic step-by-step procedure and constant overt knowledge of success is likely to become more and more important in education. Whether language teaching makes much use of teaching machines or of programmed textbooks it will always have the advantage over other teaching of being able to make use of the language laboratory. The incremental step-by-step orientation nature of such teaching learning is central to P.L., whatever the method or device used (and teaching machine, textbook and language laboratory are all equally devices); so is the validity of the teaching learning in the sense of its efficiency in teaching, i.e. it has been tried out sufficiently to show that it *does* work. The step-by-step procedure may involve a reinforcement by the presentation of correct responses (in a special frame, on a certain page, etc). Doubts have been cast on the need actually to *show* such answers. What is perfectly clear is that whether responses are shown or not the reinforcement (the *knowledge of success*) is built into a good programme by its use of the content analysis (into *steps*) and by the experimental try out and consequent item analysis.

Thus the content of what is to be taught and its representation have been incorporated into a form of self-tuition. In this sense progress in learning is being tested from step to step so that while learning is ensured (i.e. mastery of each step) incremental learning is tested at each step. Each step is, from this point of view, both teaching and testing at the same time. Mastery over the *content* of a program is thus insured by going through it: there is no need to test *after* for learning. It may, of course, be necessary to test for *memory*. It may also be useful to get students to write in responses to each frame, not for their own learning so much as to demonstrate that they have read (or worked at) it right through.

What is being suggested here is that P.L. has radical implications for testing, that, since it incorporates a kind of fail-safe device the need for after testing (what will be called Achievement Testing 3.3.1) decreases. This is, of course, admirable. Testing is not an end in itself: its essential aim is to provide useful information which will enable meaningful decisions about future action to be taken. The more that testing and teaching can be integrated and the less the ostensible demands of the external test/exam, the better for learning. P.L. at the moment seems to promise something like such an integration, and this may, in fact, be its major contribution to education.

3. TYPES OF TEST I: USES

3.1 *Language as A Special Ability*

The notions of general ability and group abilities are not so fervently proclaimed by psychologists these days. However, there is much to be said for such notions since they tend to agree with common sense observation: thus general ability and intelligence; group abilities and Ved. or Mech. and special abilities and, say music. Now language enters into very many activities and it may well be seen best as a group ability (some would see it as equivalent in status to g) or a special ability.

The fact is, of course, that where it does enter into other abilities it may well be indistinguishable from intelligence. For present purposes what matters is language as a special ability since we are concerned with various types of control over language (especially 2nd languages) and how to evaluate them. Naturally there is no clear-cut division between language as a special ability and language as it enters into other abilities and so it must be remembered here that although we speak of language as a *special* ability no assumption is made about its isolation from other abilities and activities. It is, often enough, only in the other activities that language can be observed: as a skill it may be said to exist for these other activities. But we call it a special ability because *as far as possible* we wish to separate it from those activities in which it takes part.

3.2 *The Taxonomy of Use*

Language tests may be distinguished in different ways. We shall here use three taxonomies and the first is that of *use*. As will be plain the distinctions between terms and the meanings given here are by no means universal: this is a pity but fairly common in language work. What matters, of course, is that those distinctions that seem necessary should be made: the terms applied to them are neither here nor there though it should be noted that e.g. aptitude here may be used for a use which other writers may term proficiency. Four uses may be distinguished: these uses, since they refer to types ov control may, with some slight assurance, be termed in themselves special abilities. These are: Achievement, Proficiency, Aptitude, Diagnostic.

3.3 *The Four Uses*

The four uses are distinguished first in terms of time, of past and future interest and control, and second in terms of subject matter. Thus, if we use an arrow of time going both ways, and if X is the point at which the test is taken, then we may symbolise as in Fig. 1.

Achievement : ← X
Proficiency : ←----X → Y
Aptitude : (X) → X
Diagnostic : ← X →

Fig. 1

3.3.1 *Achievement*

Achievement (or Attainment) Tests are concerned with assessing what has been learnt of a known syllabus. This may be within a school or within a total educational system: thus the typical external school examinations ('O' level, 'A' level, Highers in Scotland), the university degree exams and so on are all examples of Achievement Tests. The use being made of the measure is to find out just how much has been learnt of what has been taught (i.e. of the syllabus). Achievement type tests end there, of course: although the primary interest is in the past, i.e. what has been learnt, very often some further use is made of the same test in order to make *meaningful decisions* about the pupil's future. It would, presumably, be possible to be interested entirely in the past of the pupils: CARROLL's *meaningful decisions* then would refer to the syllabus, i.e. to any necessary alterations to it that might be necessary or to the teaching method etc., for the next group of students. But, as is obvious, Achievement Tests are almost always used for other purposes as well: it is important to recognise this and to account for it in one's test construction. All an Achievement type test can do is to indicate how much of a syllabus has been learned; it cannot make predictions, of itself, as to pupils' future performance unless the *syllabus* has been deliberately (as it should be, of course) designed for this purpose.

3.3.2 *Proficiency*

Proficiency Tests, as we see it, are concerned with assessing what has been learnt of a known or an unknown syllabus (hence the broken arrow in Fig. 1). Here we see the distinction between Proficiency and Achievement. In the non-language field we might consider, say, the Advanced Motorist's Test as a kind of Proficiency test since there is the desire to apply a common standard to all who present themselves whatever their previous driving experience: over which, of course, there has been no control at all. In the language field there are several well known Proficiency Tests of the same journeyman kind: the *Cambridge Proficiency Exams, the Michigan, TOEFL and EPTB*. These all imply that a common standard is being applied to all comers. More sophisticated Proficiency Tests (sophisticated in *use* not in design, of course) may be constructed for the all important purpose of deciding just how much control over a language is needed for certain purposes (e.g. university studies in a 2nd language). The Proficiency Tests just mentioned may well have started off as research measures to fulfil exactly such a purpose: it may be hoped that they did so, for otherwise it is difficult to understand what type of control they are intended to represent.

If Proficiency Tests differ from Achievement by the uncertainty as to previous instruction, it is in their relation to future needs or control purposes that they are distinguished from Aptitude Tests. Some writers do not make this distinction. It seems to us a useful one and one we shall elaborate in the next section under the Aptitude use. Here we will only say that Proficiency in a language implies control over adequate skill (i.e. language) for an extra linguistic purpose.

3.3.3 *Aptitude*

Aptitude Tests assess proficiency in language for language use. This is their distinction from Proficiency Tests. Hence, in Fig. 1 the arrow relates the language test under Proficiency to *something else* (Y), and under Aptitude to itself (X). Thus, as we have suggested above, we can think of a Proficiency Test as assessing adequacy of control in a 2nd language for studying other things than that language through its medium. An Aptitude Test is generally conceived as assessing amount of linguistic skill in itself for the purpose of learning languages. Generally it will mean for learning a particular language. A number of problems, of course, arise: is it just linguistic skill: if so, what is it? It is just here that considerable research is needed. Of course it is not only *just linguistic skill*: what an Aptitude Test should incorporate is that area of skill which may be termed linguistic and which is separate from the other obvious factors such as intelligence and motivation - and the whole teaching situation. It may be that once these factors are accounted for then nothing is left, hence the need for research.

In Fig. 1 we have represented the Aptitude Test X as (X), i.e. in brackets. We know where the arrow leads - to Proficiency (X); and so Aptitude research must set up its goals first, must decide on what is meant by particular language Proficiency or accept ready made ones which are commonly used in an educational system (e.g. external school exams, 'O' level and the rest). This Proficiency measure is the criteria. But what is not known is how this may be predicted, i.e. how success in learning, say, Chinese, may be predicted before any Chinese has been learnt. Naturally, except with very young children, there is always the mother-tongue to fall back on. But it does seem to us that even if the mother-tongue is (as it almost certainly has to be) used for predicting success in foreign language learning, here is an important area for research, inasmuch as it will serve to break down into component parts those skills in areas in present linguistic control which relate to future control. If in fact this can be done then it seems immaterial that it is being reached through the medium of the mother-tongue. The importance of work on Aptitude is not so much to predict future successful learners as to indicate which is fundamentally involved in language learning.

The distinction we have drawn between Proficiency and Aptitude may not always be a real one. Is a Pilot's Proficiency Test really an Aptitude test? Such a Proficiency Test could be looked on as a battery of Aptitude Tests (e.g. manual dexterity, sensitivity to lights, etc.) which together have been shown to predict performance as a pilot. But in language terms it seems sensible to make this distinction and to think of the Proficiency Test as facing both ways, as taking place at some point *after* learning has started, and as relating to future non-language performance. An Aptitude Test, in these terms, is concerned with inherent aptitude for language learning.

3.3.4 *Diagnostic*

The Diagnostic one is an odd man out because it relates entirely to the use of the information (which, of course, all the other uses do, too) and not at all to the presence of a skill in the learner. Achievement, Proficiency and Aptitude are all concerned with both *use* and *skill*. A Diagnostic Test is a use made by a teacher etc., of the information provided by one of these skills. A Diagnostic Test may be constructed for itself or it may be an additional use made of an Achievement or Proficiency Test. If it is specially constructed it could perhaps be argued that some element of learner's skill does enter in, or rather absence of skill because the tester is concerned with discovering what might be termed Non-Achievement.

Like the Proficiency Test a Diagnostic Test looks before and after (or it may so do): before to find out what is wrong with the previous learning, after in order to *do* something about it, i.e. for purposes of planning etc. remedial work.

Non-Achievement may well be a sensible name for a Diagnostic Test. Just as the Achievement Test is built on the content of the previous learning, i.e. a sampling of the syllabus that has been taught, so the Diagnostic Test is built out of what has not been learnt: it may be exploratory in that it sets out, based either on syllabus content like an Achievement Test, or on the content of some language description which indicates certain points of contrast or interference between mother tongue or first language (L1) and second language (L2). In point of fact if the language description is satisfactory the syllabus content will certainly relate to it, i.e. will be partially based on it itself.

Once such exploratory work has been done and the likely errors uncovered, future diagnostic tests can be, at least partly, made up out of *expected* errors. We say *at least partly* because there is a danger of such Error Analysis work in that it concentrates on language interference and ignores learning problems of other kinds (An approach based partly on syllabus content - where the syllabus is related to future needs as well as to interference - will avoid this danger). When the *whole* of language learning is reduced to interference, i.e. semantically and contextually as well as formally, then this is at least logical: what it says is that *every* language learning problem can be related to the L1 of the learner. If the non-formal as well as the formal features are accounted for then there must be some sense in which this is so. But there is a tendentious ring about it.

3.3.5 *Summary*

We have been discussing some of the *uses* of language tests. They can be distinguished, as we have suggested, in two main ways, in their connection with a known syllabus and in their relation to the time scale. The one affects the other: Achievement (at the one end) has a known syllabus and concerns the past; Aptitude (at the other) has no past (except, of course, that it must relate to the learner's previous experience, but that is not the point here) and concerns the future. Proficiency and Diagnostic both concern past and may or may not have a known syllabus: both relate to the future. Proficiency (unlike Aptitude) to a non language skill; Diagnostic (like Aptitude) to a language performance.

But if these are distinct *uses* they do not necessarily refer to distinct tests: the same test could very well be used for Achievement and Diagnosis: if it is based on a *favoured* syllabus then it could be used for Proficiency also (*favoured* in the sense of a proven satisfactory course which not all courses will necessarily have followed: this is doubtful, of course, because what starts off as Proficiency quickly ends up as Achievement simply because discrete learners adapt in order to make themselves more suitable for their *Proficiency* Test which to them is an *Achievement*). If this all purpose test is given very early on in a language course, say at the end of 3 or 6 months in a school course then it could also be used for Aptitude purposes if any predictions are made as to likely future language learning success. A *use* then is not a *test*: it may (as Aptitude almost always will) need a separate instrument, but not necessarily so. What matters is that the user should be fully aware of his purposes.

4. TYPES OF TEST II: EXAMINATIONS AND TESTS

4.1 *Test/Exam Distinction*

Another way of looking at language tests is to make some attempt at resolving the test/exam distinction which is rarely made. How does a test differ (if at all) from an examination? The notion of test conjures up vage ideas of psychology and of intelligence whereas an exam suggests the end of term multi-subject circus. Is a test more objective, more reliable, more specific, more realistic than an exam? We shall try to show that it is all these but that an exam may have an advantage in its influence over the curriculum. We shall also suggest that while these distinctions exist for the traditional test/exam difference there is no reason at all why an exam should not build into itself whatever virtues a test may have. But the traditional distinctions remain.

In common usage the terms are used synonymously to a certain extent. The famous 11+ test is sometimes called an Exam; and all school children are well acquainted with class tests which occur once every week or fortnight and so on. Typically the Examination is a more full dress affair than the Test: the 'O' level or end of year test in school, the degree exam: these are examinations. Also the Examination is, in our terms, usually concerned with Achievement: not always though - the university scholarship *examination* is more a proficiency or even an aptitude type test. And our kind of Achievement Test is certainly not most clearly seen as an Examination. There are other distinctions.

One simple but probably unacceptable solution would be to make the term *Examination* refer to the whole area of language measurement and the term Test to a specialised part within it, i.e. a *Test* to be seen as a kind of exam. But since this is not how the terms are used and since specific examples of the two types are sometimes so different that they cannot be drawn together, what we shall do is to look at four levels, as it were, and suggest ways in which exams and tests are drawing nearer.

4.2 *Exam/Test Levels*

We are suggesting four levels: no suggestion is made as to hierarchy:

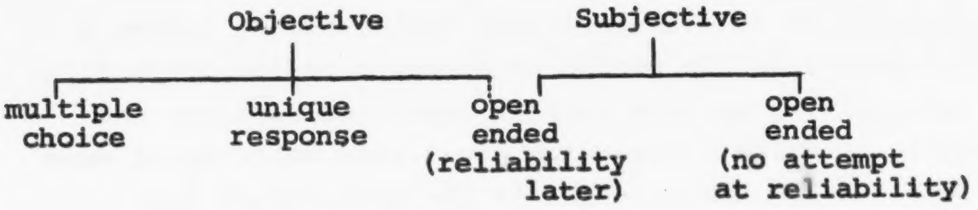
1. Examinations (traditional)
2. Ad hoc tests (teacher made)
3. Standardised tests
4. Examinations (new type)

We shall make a start with standardised tests since it is in relation to them that we wish to consider the others, and it is because of them that much of the testing mystique has arisen.

4.3 *Standardised Tests*

A standardised test is objective; by this is meant that it has been tried out on a proper sample of the population for whom it is intended and that on this sample it has been shown to work, i.e. that it is both reliable and valid: it is made up of *items* each of which is discrete in itself and has been shown to contribute towards the total performance of the test. As well as being objective a standardised language test will rest on some adequate linguistic basis, i.e. according to its use, it will be made up of items which are linguistically meaningful. So a standardised language test has two qualities: it is psychometrically objective and it is linguistically realistic. Of course it may be argued that the second is subsumed in the first since under objectivity we have already mentioned *validity* and presumably by validity is meant *linguistic* validity of some kind. This is so of course and in the Achievement Test situation objectivity would be enough since the syllabus is (we hope) linguistically realistic. We are thinking more now of Proficiency (and perhaps Aptitude) Tests. Here the objective validity (and psychometric quality) relates to the linguistic section (itself embodying an acceptable theory of language). That is to say that in the Proficiency situation the Proficiency Test must, as it were, assume its own syllabus and also make itself objective on the basis of this syllabus.

There is a further problem about *objectivity*: if a test is objective does this mean that all the items must be objective too? Does this mean that the Essay, for example, cannot be used in an objective test. All items, since all must contribute to the test, must be objective. But this does not necessarily mean that the objectivity must be built in in the form of multiple-choice answers. It usually will because it is easier and because the analysis of trial items (Item Analysis) will have shown whether they work. But the Essay type question is not precluded. This is best seen as a type of open-ended question: it is possible to use these and then attempt objectivity in their scoring. This is more difficult than the multiple choice method: it introduces a second type of reliability problem, that of Examiner unreliability. Still this can be accounted for and several experiments have indicated that Essay unreliability can be taken care of. A diagram may show the connections:


Objective
Subjective
multiple choice
unique response
open ended (reliability later)
open ended (no attempt at reliability)

The problem of a language proficiency test which is to be widely used in ecucation is that reliability and what most teacher users would regard as validity are inextricably intertwined. That is to say that as reliability is raised (by objectivising through use of multiple choice and so on) so validity is lowered (because of the influence on the classroom) and this for two reasons. First some skills, which cannot be *reliably* tested, are very often bypassed: e.g. Oral Production, Continuous Writing (e.g. Essay). Second, the influence of the objective item test is to force the teaching which necessarily precedes the test into practising such items. As has been agreed above this becomes a problem when a Proficiency Test is looked on as an Achievement Test. But if standardised test techniques are introduced into educational examining then it is essential to prevent the deterioration of real validity, i.e. to ensure that all the meaningful skills are included and that the teaching does not become item practice of another sort. One argument against the traditional exam is that it gave no help other than itself as a syllabus model: if standardised tests influence the new type exam they must do much more than this. But the danger is there.

To say that a test is valid is no answer to the critics of this kind who are really saying that they don't like the educational aims that such validity imposes on them.

4.4 *Ad Hoc Tests*

These are quickly dealt with. All that is meant by ad hoc tests is those tests which any teacher may wish to construct and use himself in his own teaching. Since he is generally concerned with a whole population and not with just a sample (i.e. his population = his class; he is not concerned to generalise from them) he is not concerned so much with item analysis: but he is concerned with reliability and validity. The latter he will know about: he will obviously be concerned to test what he has taught and will not make use of any techniques of testing (e.g. *multiple choice*) unless he can cope with their back-wash in the teaching. But he should also take reliability into account. The more, therefore, the test approaches the nature of a standardised test (while, of course, remaining under his control) the better for it. What ad hoc tests can learn from standardised tests are, first, the importance of reliability, second the value of item discreteness; and third the importance of thorough linguistic analysis of the content to be sampled for the test.

Teachers have always made use of such tests. They have always, it is important to make clear, made use of versions within their own classrooms of standardised tests. Such tests may be more readily accepted if teachers agree that they are nothing more than generalisations across the experience of a large number of teachers.

4.5 *Examinations (Traditional)*

The typical F.L. examination (old type) consisted of various types of exercise such as comprehension of passages in the foreign language, translation from (and sometimes into) the foreign language, an essay in the foreign language. These were, of course, all reading/writing tests. There might also be an 'Oral' exam of some kind given either by a visiting examiner or by the candidates' own teacher.

What is notable about such an exam from the testing point of view (and in terms of the criteria we have been discussing) is not that it lacks initial objectivity (all exams do) but that it lacks linguistic realism in terms of current attitudes to language learning. If languages are learnt in order to be spoken (though of course they are also learnt to be read) then the teaching must emphasise this. But so it does: or is supposed to do. It is the exam (traditional) which is out of line therefore in that it emphasises skills of a kind which were once thought pre-eminent but now considered less so. There are, of course, other kinds of doubts about translation and about essay writing too but of a different nature. That is to say that there are three levels of approach: first that the purpose of language learning has changed from entirely reading to mainly speaking; second that translation *in any case* (whether the purpose has changed or not) is a specialised kind of exercise and perhaps best not tested in this way; third that essay writing needs special care in order to be reliable as a test of the written language. There are all kinds of validity argument. What must be stressed here is that the traditional exam was highly valid as a test of what was regarded (and no longer is) as the purpose of language teaching.

If it is, as it could be, maintained that such reading/writing tests are more *reliable* than more recent speaking/listening ones then it would, we feel, be in order to bring up the argument against such a position that we described earlier (under standard tests) in the tension between reliability and validity. However reliable a test it is no use if its validity is not what is wanted from it.

4.6 *Exams (New Type)*

This distinction between Exams traditional and new type is, of course, an artificial one. Two such distinct types do not exist and no one sets out to construct a traditional exam. However, there are signs of a new kind of exam emerging and it may well be that its constructors do in fact think in terms of constructing a new kind of examination. Such an exam (new type) is characterised first by its approximation to the Standardised Test, i.e. by its attempt to build (psychometric) objectivity and (linguistic) realism; and second by its attempt to work in detailed closeness to a syllabus. This second feature is the clear answer to the problem of the Proficiency type test having to assume its own syllabus for which it then becomes a kind of Achievement Test. What is needed is for aim, syllabus and exam all to be blended in together, for exam to relate to the content of the agreed syllabus so that, in fact, it is possible for a test - like Exam to retain its Achievement use. Apart from this most important feature the special characteristics of the new type exam are seen in its attempts to incorporate testing strategy. An example may make this more clear.

Traditional type	New type
1. Aural Comprehension 5 q. open-ended	1. Oral Reading
2. Read. Comprehension 7 q.E open-ended	2. Responses to Questions in f. 12 open-ended
3. Trans. F ⟶ E	3. Oral Comprehension-Pictures open-ended
4. Read. Comprehension 12 q.F open-ended	4. Aural Comprehension - 4 voices open-ended
5. F Essay	5. Read. Comprehension

Fig. 2

As Fig. 2 shows the Exam (new type) abandons the Essay and the Translation: it omits an Oral interview. It retains only one sub-test similar to one in the traditional exam, the Reading Comprehension. The other four are all concerned with the Spoken Language and all are recorded on tape for later scoring: the first is a general assessment of reading aloud; the second an attempt to simulate a question -answer situation in which candidates must respond in French to questions in French; the third provides an opportunity for structural oral composition by means of pictures as stimuli; the fourth simulates a conversation situation with several voices after which the candidate must show that he has comprehended what he has heard.

The characteristics, we have suggested, of Exams (new type) is that they bring in testing strategy (objectivity and realism) and that they align themselves to a Syllabus (which is another way of saying that they demonstrate in detail their linguistic realism). So far as is known neither of the exam types set out in Fig. 2 has a detailed syllabus associated with it. However, this is more a lack in the traditional type than in the new type. Both should have a syllabus: but what the new type has done is at the least to look at the needs of the pupil and the ways in which the schools are setting out to meet them, has attempted, in fact, to be linguistically realistic. The traditional exam was 80% reading/writing; the new type is 80% speaking/listening. Why? Because it is assumed now that these skills are more suitable for the learner to control; and, in fact, it is on these, more and more, that the schools concentrate. Not entirely; and not all schools (the exam remains an *alternative*). The new type may, therefore, not provide a syllabus; but it does look at what seems to be necessary. We would claim, therefore, that it is linguistically realistic and, therefore, beneficial in its backwash on the educational system.

But what of its objectivity? If it is linguistically realistic then it must have some objectivity, i.e. it must be at least partially *valid*. All its items are open-ended: objectivity of the *reliability* kind needs to be built in after the test has been taken. Each tape was heard (at *first*) by two or more trained examiners in the pilot experiment. Later on only one examiner scored each tape: by then, of course, they had had plenty of practice. But there are two sources of remaining unreliability: the usual examiner one (of the written essay kind) and the more specifically spoken language one where different examiners (unless they are experienced phoneticians) cannot agree on what they have heard nor what response is expected of them. Either may be allowed for: together they pose serious and enduring problems. It is hard to find fault with so serious an attempt to meet all demands. The only one we would make (on the theoretical side) is that of doubtful Reliability: this, as we have suggested may be unavoidable in order to retain Validity. But there seems no reason why at least some of the items should not be constructed with multiple choice responses. Even half constructed on this basis would go a long way to meeting this problem of reliability without interfering with Validity through washback in the schools.

This particular new type exam was widely tried out experimentally and its items and mark scheme altered accordingly. A statistical analysis of the results was made. Here again there is a serious attempt at objectivity. All in all the new type exam goes a long way to answer language learning needs and to meet the criticism of the traditional exam.

5. TYPES OF TEST III: LANGUAGE LEVELS AND SKILLS

The third division into *Kinds* of language tests is that of *language levels and skills*. This is the most common approach. Thus we speak of Oral tests, of Listening tests, of anything which is itself simple because it isn't everything else. Other names which enter into this category are Grammar, Comprehension and Reading Speed. Naming of exercises which is what these labels are is another way then of considering types of test. Most language teachers are concerned either with proficiency or with achievement uses; their concern is with *what* needs to be tested and by *what* they are thinking of both a language level (e.g. grammar) and a language skill (e.g. reading). They are anxious to know whether they should test reading, speaking, dictation, précis, essay, vocabulary, comprehension, spelling and so on. How, they are asking, can testing techniques help them get at skills in language.

A distinction needs to be drawn between testing techniques and language skills. Reading and speaking, comprehension and writing are obviously kinds of skills. On the other hand dictation, précis, essay, speaking: these are testing devices for getting at language skills and therefore appropriate if they are considered to provide useful information. The trouble with the testing devices is that they tend to be umbrella type tests, not analysing sufficiently so that a mistake in say précis cannot easily be related to a skill (is it reading or writing and so on) and that they take the aura of language goals and thus become taught as if they were themselves skills. The most economical way of balancing skills against levels is to draw up a plan with the skills of reading, writing, speaking and listening along the top and with the levels of phonology, grammar, vocabulary and context down the side and at the meeting places of each pair to fill in a language testing technique. Thus where reading and phonology meet will appear tests of phoneme discrimination, of awareness of stress and intonation; where reading and grammar meet will appear tests of grammatical choices; where reading and context meet will appear reading comprehension. The purpose of such a plan as will be obvious is of course to ensure that nothing is omitted. It cannot be stressed too much that the *technique* itself is trivial in the sense that it doesn't matter how it is done as long as it provides the information required. What matters very much is the tester's view of language and of just what is a skill. This will influence his choice of a plan or model on which to base his test.

MEASUREMENT OF ORAL COMMUNICATION

John A. Upshur

It is a great pleasure for me to be here addressing this group today. That sentence, which I've just uttered, is well-formed - "correct", if you will. But one view would have it that it was nontheless *not* real language. It is a sentence used before at many times, in many places, before many audiences, by many speakers. It is not something that I have created, so although it has the outward form of language, it is not language. It is more of an involuntary flinch produced by the sight of an audience of respected and critical colleagues.

The rest of my remarks will be made in real language - the sentences will be novel. I trust that at least some of the facts, opinion, conjecture, and orientation will be new as well. In order that I do not spend your time repeating what you already know, I will not attempt a lengthy documentation of the lack of success in testing oral production. Instead, I will briefly indicate some of the works on oral production testing and suggest some of the reasons offered to explain why they have not satisfied demands. Then I will suggest a model and techniques for testing, and finally describe some of the preliminary work done at the *University of Michigan*.

I Background

In recent years educators have been not only continuing, but even increasing their calls for *production tests* in foreign languages. It may be rightly argued that foreign language tests of almost any sort require of students processes fundamental to linguistic production, hence that production ability is being measured even though the measures are indirect and only roughly approximate. It is also true that existing test batteries do include compositions, interviews, oral readings, picture descriptions, mimicry and substitution tasks, and question asking, but the clamor continues.

LADO (1961) and VALETTE have suggested a number of ways in which speech can be tested. RAND (1963) has described a short test of oral English proficiency, and NEWMARK and SWEIGERT have described in detail criterion referenced oral production tests for elementary language courses. JAKOBOVITS (1968c) has presented a summary of techniques for testing bilinguals which includes nine methods involving actual production. OSSER (1968) has illustrated and summarized techniques to test for acquisition of specific grammatical rules, and MCNEILL (1966) has demonstrated a non-metric measure (grammar writing) of language ability based upon the oral production of first language learners. Presumably this method would apply also to the characterization of second language learners. This does not begin to exhaust the list, yet still the cries for production tests go forth. It would be comforting to conclude that foreign language educators do not read what has been written for them, or that in their ignorance of the occult ways of the *measurement fraternity,* they irrationally insist upon neat published packages with high face validity - that is, that they make unreasonable demands. But such comfortable conclusions are unwarranted. A closer look at some of the less public explanations by the testers shows that this is so.

One of the reasons offered to educators for the lack of the kind of production tests they want is the fact that a *general proficiency factor* accounts for much of the variance which might be expected in such tests. Yet the *Test of English as a Foreign Language* still contains five sub-tests even though it is almost entirely a test of two highly correlated factors. This suggests that, at best, there is disagreement about the extent to which *a general proficiency factor* is included in *production ability*; that, at worst, the general factor argument is an argument of convenience.

A second reason is the assumed high correlations in the four skill model of language testing (cf. CARROLL, 1961) between production skills and corresponding reception skills. Although PERREN (1967) has cautioned correctly that there is still too little empirical evidence to support this assumption, a number of pertinent studies have been made. PITCHER, RA, HARRIS, and PALMER (1967) report a correlation of .65 between ratings of written themes and the *TOEFL Reading Comprehension* sub-test. The theme ratings correlated higher, however, .74, with an indirect measure of writing ability than with the corresponding receptive skill, reading. Concerning measures of production which do not require actual production, UPSHUR (1962) found that scores for students of English as a second language on a multiple-choice test of English word-order correlated .86 with scores on a test in which they were required to arrange the same words in the same sentences for themselves; but the correlation (C) between responses for items on the two tests was .43. PALMER and UPSHUR (1960) found a correlation of .86 between scores on a test of the ability of second language speakers to recognize rhymes in a series of printed words (e.g.: *fox, locks, cots, logs*) and scores of the number of keyed rhymes that were actually made in reading the lists. And a further analysis of data yielded a contingency coefficient of .15. The total number of written errors was approximately half the number of items containing speech errors, and even so, only about half of the written errors could be attributed to speech errors. Two conclusions from these three studies appear tenable: first that indirect measures of production are more accurate (i.e., correlate more highly with direct measures of production) than are corresponding reception measures; second, that indirect measures not requiring actual production may fail to provide desired and useful information.

Another reason given for the dearth of production tests may be more directly attributable to the influence of the 1961 CARROLL model. The model seems to represent a hasty marriage between a notion about language skills enunciated by PALMER (1921) some twenty years earlier, and structuralist notions about *linguistic aspects* which were, if perhaps on the wane, at least current. Viewed today (the same view might have been taken much earlier), the model includes little or nothing in the skill domain which may not be considered a proper part of linguistic competence. The test writer schooled in CHOMSKY's (1965) distinction between *competence* and *performance* will likely find little in empty cells of this model to draw his attention to aspects or processes of a human performance component of language behaviour. Even the test writer who shares BOLINGER's (1968) apparent reservations about the utility of the competence-performance distinction finds little to guide him. COOPER (1968) has already pointed out some of the severe limitations of this model. Finding others, simply for the sake of announcing them, does not, however, appear to be an especially useful task.

Many other ostensible reasons are offered, but I will mention only one more here. It has at its base a language acquisition theory which strongly emphasizes habit formation. The habit theory, when focusing upon sequential dependencies among surface forms, tends to promote *pattern-practice* testing. But JAKOBOVITS (1968a) has shown the theoretical inadequacies of *pattern-practice* learning, and BRIERE (1968) has documented the *patternese* responses (not necessarily indicative of speaking ability) to *pattern practice* testing. The habit theory when focusing upon situation-as-stimulus and utterance-as-response tends to promote dialog-memory teaching and dialog-recitation testing. Yet SPOLSKY (1968) has consistently argued the inherent limitations of judgements based upon language-like behaviour (such as dialog recitations) and the need for tests of *creative use* of language.

It is easy to refute the arguments which are offered to show that new production tests are unnecessary. This suggests rather strongly that such tests *are* needed. I do not intend to imply, however, that no responsible test writers have recognized this need, or that no one has given sound reasons why the need is still unmet. I shall take the position here - without arguing it - that the need has not been met primarily because of the lack of any performance theory generally available to, and useful for, those who might prepare new production tests. This is not to deny that language performance models have been constructed. CHOMSKY's (1959) scathing review notwithstanding, SKINNER's (1957) model has some explanatory power for the learning and operation of what OSGOOD (1967, p. 113) has termed "sequential hierarchies" in language performance. More recently model builders have been most concerned with theories of language users which MILLER (1964, p. 94) characterizes as describing the processes whereby the language user succeeds or fails in obeying the abstract rules of grammar. PERREN (1968, p.112) suggested demands that a model useful for language testing must meet, when he stated that:

> *...overall competence in spoken language must be judged ultimately in terms of intelligibility and communication. But intelligibility is not merely the production of identifiable sounds in a conventional order;it depends on a situation which includes a particular listener. The meaning of what is said must be the final criterion, and it can only be judged by reference to a context of situation and audience.*

OSGOOD supplied two hints about the way a useful model of a language user might operate; first when he quoted MOWRER's notion of communication as a process of transferring meaning from sign to sign within a person rather than transferring meaning from person to person (p. 105); and second when he noted (p. 125) that *the speaker is never ambiguating ...and the hearer is never ambiguated in ordinary discourse.* These ideas are reflected in a rough model which has been useful to me as a guide to thinking about production tests.

II A Communicator Model

It is not my purpose here to present in detail an extended model of the language user. Rather I intend to sketch in some of the components of such a model which might take into account what GRICE (1968) has called utterer's meaning, that is (p. 255):...*that notion which is involved in saying of someone that by (when) [saying 'such-and-such'] he meant that so-and-so..., [as opposed to] the notions of meaning involved in saying (1) that a given sentence means 'so-and-so' (2) that a given word or phrase means 'so-and-so'.*

This sense of meaning I shall call *Speaker's Meaning (SM)* as distinguished from *Utterance Meaning (UM)* or *Word Meaning (WM)*. In communication it is the task of the producer (a task certainly shared by the receiver or audience) to *induce* in, or transmit to, the audience a meaning, an *Audience Meaning* which has as a part an equivalent of the *speaker's meaning*. For example, the sentence, *What a delightful twenty-hour plane trip I had*, has as one utterance meaning: *I enjoyed the long flight*. The speaker's meaning might, however, be that he was exhausted. Communication has taken place if the audience decides on what the speaker actually meant rather than on another possible utterance meaning; communication has also taken place if the audience derives a meaning, part of which is equivalent to the speaker's meaning. For example the meaning he derives might be that, *Speaker thought that he was exhausted*, or *Speaker wants me to believe that he was exhausted.*

Because communication requires that audience meaning contain speaker meaning, because utterance meaning is *a* medium through which this is accomplished, and because speaker meaning need not be equivalent to utterance meaning, more is required than that Speaker (and Audience) have competence in some language and that both have the ability to make a *sufficiently small* number of errors in applying the rules of their competence components. I would suggest, in fact, that not only are competence and utilization-of-competence *not* the only components necessary in a theory of communicators, but also that other components (or *factors* or *processes* - about which I will say more later) may compensate in large measure for limited competence and/or utilization-of-competence.

Language teachers supply ample anecdotal evidenc not only that there are students who can perform beautifully on substitution drills, transformation drills, etc., yet with whom communication is virtually impossible; but also that there are students who *do miserably on your tests, but we can talk about anything together*. Thus I would argue that well-formedness of his utterances is not a sufficient condition for a speaker to communicate, nor is it a necessary condition - only partly, however, because of natural language reduncancy and because an audience might possess what OSGOOD (p.117) terms a sensory integration system.

At this point it seems necessary to present some notion of the term *meaning* especially as I have used it in *Speaker Meaning* and *Audience Meaning*. Two useful points of departure are (i) FILLMORE's (1968) ideas about case relationships and about proposition and modality constituents of the base of sentences, and (ii) QUILLIAN's (1967) view of semantic net structure. These ideas seem to be useful approaches for constructing a theory which will include the three levels of meaning proposed by JAKOBOVITS (1968c). The first level, which he calls *linguistic meaning*, is the traditional concern of linguists. I have referred to this as utterance meaning. This abstract meaning is derivable by utilization of linguistic competence: phonological actualization rules, a dictionary of lexical meanings and their projection rules, and a *theory* defining syntactic relations.

JAKOBOVITS uses *implicit meaning* to *refer to the eliptically derived conceptual event which an utterance represents*.
He illustrates the meaning of *agree* in a number of sentences such as:

(a) The strikers agreed to go back to work.

(b) Mary agreed to marry him.

(c) The university agreed to pay his moving expenses.

In each sentence the verb has some different and particular meaning which can be inferred from the situation to which the sentence as a whole refers. Dictionary meanings of words apply restrictions upon the particular meanings they may have, but within the bounds of these restrictions, meanings are potentially infinite. JAKOBOVITS argues that what the speaker intends in communication is the particular and not the general or abstract.

At a higher level, JAKOBOVITS proposes *implicative meaning*. This includes intentionally or unintentionally communicated relations between the speaker and his utterance. Among these are his intention, claims about the status of his utterance, and even such things as his geographical origin.

FILLMORE has characterized a sentence (S) as composed of a propositional (P) and a modality (M) constituent:

(1) Sentence $\longrightarrow$ Proposition + Modality

The propositional constituent is made up of a verb (V) plus one or more case categories (C):

(2) $P \longrightarrow V + C_1 + C_2 + \ldots + C_n$.

Speaker meaning of a sentence may be viewed in similar terms as a concept including a propositional constituent, and also including relations between the proposition and other concepts, for example, temporal relation to time of communication act or to some other time of some other proposition, compatibility relations to other concepts, etc. Thus FILLMORE's characterization of a sentence may be transformed into a concept characterization of the form:

(3) Concept Proposition + external relations...

His *rewrite* of the propositional constituent of a sentence becomes: (4) Proposition $W_V + W_{C_1} + W_{C_2} + \ldots + W_{C_n}$ in which *words* are the *entities* having case relations in the proposition.

QUILLIAN's semantic net theory is powerful enough to account for potentially infinite word meanings - a necessary condition for an elliptically derived conceptual event, that is, the appropriate meaning created by the audience when communication has been successful. In a semantic net, units are not pictures, words or sentences as generally conceived, but what are ordinarily called *attributes* (*word form*, however, can be an attribute in a concept). A word concept involves a vast amount of information; it represents relationships between memory elements at least as varied and complex as any which can be expressed in a natural language. But the information associated with a word concept is differentially accessible depending, for example, upon the path(s) in his semantic net an audience takes after he enters his store at the *word form* node, or the path a speaker took which included the same *word form*. A word in a concept is the information contained in the restricted or curtailed path of a QUILLIAN type semantic net. A crucial advantage for this type of theory is that, because units are attributes, an audience has a means of access to semantic information by other means than the word forms of a speaker's utterance.

To summarize briefly, in order for a speaker to get his meaning across (that is, to communicate, to have his audience create a meaning which contains the speaker's meaning), it is necessary (1) for the audience to know the case relations for each word in an utterance (this *may* be accomplished by utilization of competence); it is necessary (2) for him to know word meanings, where word meaning refers to a restricted path through a net (not to *word concept* which includes all possible paths); it is necessary (3) for him to know the relations between a proposition and other concepts. A performance theory which explains only the successes and failures of a Speaker (or Audience) to apply his competence may account for communication of case relations; it can perhaps (although I doubt it) account for the communication of word meaning; but it cannot altogether account for the communication of relations between propositions and other concepts. In other words, the modality constituent in a sentence theory does not include all of the relations between a proposition and other concepts which are important for a theory communication.

Before outlining the operation of a model of communicators, I would like to note an implicit change in the meaning of the word *creativity* as it has been used in this paper. Initially I spoke of linguistic creativity as roughly defined in terms of a speaker uttering sentences different from any he has heard before. The emphasis has changed to what I might call communicative creativity. This suggests a speaker making an utterance in some situation which in effect causes his audience to create a concept which he never had before.

The leap from this idea of communicative creativity to a model of communicators is not a long one. From the Speaker's point of view, he must have a concept to communicate (Speaker Meaning, or SM) and some reason for doing so. I shall say nothing here about motivation except that a part of it depends upon a belief by the Speaker that his Audience lacks the concept SM, and cares to have it.

In this way motivation in a communication model is not necessarily the same as in a model for talking. As an experienced audience himself, a Speaker also implicitly knows what processes his Audience must carry out in order to obtain SM. His ability to communicate an intended SM is therefore constrained by his audience's linguistic competence, his ability to utilize his competence, by the perceptions available to him by his experience as reflected in his semantic net, by what has already been communicated, by his knowledge relative to *word forms* used, and by his inferential ability. Because these constraints define the components of operations of the model, let me repeat them . The ability of a speaker to communicate depends upon his audience's linguistic competence, his ability to apply the rules of his competence (*memory span, etc.*), the non-linguistic perceptions available to him, his knowledge of the world as represented by his semantic information net, by what has already been communicated during the current communication transaction, by his knowledge relative to *word forms* (that is, what information in his net he can retrieve when his access to the net is through a word form rather than some other attribute), and finally by his inferential abilities.

Depending upon a Speaker's assumptions, firm knowledge, or guesses about the contents of these components, he may have to supply prior information in order ultimately to communicate a given SM. His communication ability is then a function of (1) his success in determining the constraints imposed by the contents of his Audience's components, (2) his success in altering the contents of those components, and (3) his success (in language communication) in adapting his own competence.

I shall not here offer conjecture about whether processing is parallel or sequential, about possible sequences or about the heuristics of an inference making system. My purpose in sketching the barest outlines of a model of communicators has been to suggest ways of testing communication ability. But before turning to that question, I would like to list several other implications of the model. These can at least show us whether it is a highly unrealistic theory. In fact, it implies a number of conclusions which are, in a sense, quite trivial. It implies, for example, that you can't communicate with someone very effectively if you don't take into account what he already knows. But good teachers have certainly been saying this for a long time. It implies also that there are severe limitations on your ability to communicate with someone if he doesn't know the language you're speaking; but that the more similar the experiences and perceptions available at the time of communication between speaker and audience, the less severe the limitations. Several other such implications include:

(1) Speaking the same language doesn't assure perfect communication.

(2) The good communicator *knows his audience.*

(3) What you say today is liable to be interpreted in the light of what you said yesterday.

(4) *Saying so don't make it so.*

But that everyone knows these statements to be true does not necessarily mean that they are trivial. How much worse it would be if the model implied statements of *fact* that everyone knows are false.

The model does, however, suggest conclusions which are not so well known or widely accepted and which may be subject to investigation. With regard to the acquisition of second language competence, for example, it suggests that a learner who is motivated to communicate will develop greater competence and ability to apply competence when the second language is used to impart new information, rather than to repeat already known information. One would expect, therefore, greater competence (though perhaps not more effective communication) from learners in the more typical co-ordinate acquisition situation than in the compound acquisition situation.

There are several things that the model suggests about communication testing in general, and about oral production testing in particular. First, it *may* be possible to get measures of the various processing components of a speaker (e.g., competence, utilization of competence, etc.), but these are not equivalent to a test of communication ability; oral production testing, viewed as one of the four skill components of the CARROLL model, is but one part of speaker communication testing. It also suggests that the complexity (but not necessarily the difficulty) of a communication task for a speaker is a function of the number of his Audience's components which must be utilized by the Audience in determining SM. Perhaps the most important implication for testing, though, is that communication measurement involves a matching of Speaker meaning and Audience meaning, and *precise* measurement is not likely without comparable measures of both. It is hardly surprising that measurement of speaking ability by interview and of writing ability by *composition* typically have such low reliability. When the grader must both act as audience deriving a meaning from a speaker's utterance and also must judge whether or not an equivalent SM is *plausible*, the chances for error are great. His task requirements as audience conflict with his requirements as comparer. It does not help much if he knows SM before hearing the speaker's utterance, because this would in so many cases preclude his deriving a meaning which did not contain SM; he is left with *counting errors*.

The model implies four general types of communication situation:

1. That in which no information previously communicated during the communication transaction and no information available by means of the audience's perception is required in order for the audience to derive a meaning which contains SM;
2. That in which contents known to exist in the audience's components are required in order that a given SM may be communicated;
3. That in which certain contents are known not to exist, but necessary in order that a given SM may be communicated; and
4. That in which the contents of an audience's components are unknown, but certain contents are required in order that a given SM may be communicated.

In order to translate these communication situations into language test situations several requirements must be met. First is a naive audience to receive a signal from the speaker. Since language mediated communication is at issue the audience must be *typically* able to communicate in the test language. His naivety refers to his knowledge of SM before receiving a signal from the speakers. In some cases, however, (the second communication situation) he is supplied with certain information before receiving the speaker's test message. The second requirement is that the audience have some way of errorless communication to a grader. The third requirement is that the grader must know SM. Grading is, formally, a process of determining the extent to which SM is present in the meaning derived by the audience. In practice this process may be an all-or-none scoring procedure, an approximation to an information transmission score, or others.

Instrumentation for the first type of communication situation might assume a form somewhat like the following: grader presents S with a series of SMs to communicate; these are unrelated *facts* such as represented by:

I am 19 years old.
My aunt is a park superintendent in Riga.
The town of Picklebrine celebrates the one hundredth anniversary of its founding next Tuesday.

These may be presented to the speaker in his native language, but perhaps in an unnatural, *simple-sentence* form in order to inhibit translation across surface forms, or in a picture language like that illustrated by VALETTE (1967). The audience records the meanings he derives and the grader compares them with the speaker's intentions, the SMs. This procedure differs from commonly used methods only in employing the naive audience. The difference is a highly significant one, however.

Instrumentation for the second type of situation requires that the speaker have knowledge about prior successful communications to the audience, or that preceptions (a display, for example) are available to both speaker and audience. In the first case the speaker must know that he has successfully communicated to his audience the SMs represented by, for example,

> One little piggie went to market.
> One little piggie stayed at home.
> One little piggie had roastbeef.
> One little piggie had none.

and then must communicate the SM represented by:

> The shopping pig and the eating pig are actually the same pig, but the stay-at-home pig and the hungry pig are different pigs.

In the second case a convenient test technique is for the speaker to describe (or otherwise indicate) to a naive audience aspects of a display which are in some manner significant. The audience records his derived meanings and these are compared with SMs. Work with a test of this type has been begun and initial findings will be reported later.

For the third type of communication situation, contents of components are known not to exist, one type of test has been tried out and some others seem appropriate. In the procedure which has been tried, the speaker is presented with a task which can be described roughly as follows:

> You have a four by four matrix. In four of the sixteen cells are four different geometric figures. In each of the figures is a different letter. Your audience has an empty matrix. Your job is to get him to put the letters in the proper cells, but you may only locate the letters for him by indicating the figures surrounding them.

In other words, the speaker must first locate the geometric figures in order to then locate the letters.

For the fourth type of communication situation - that in which contents are relevant but whose existence is uncertain - one technique has been divised, has been demonstrated to be feasible, but no further work has yet been done. The speaker is given four pictures labelled A,B,C,D. He is informed that his audience has a duplicate of one of the pictures but without the label. His task is to tell his audience the proper label for his picture. As with the other test procedures discussed, the audience is naive. He indicates what he understands to be the label for his picture, and this is compared with the label on the corresponding picture of the speaker's set.

III Preliminary Findings

Several years ago I began naive audience testing in a very small way. Subjects were students of English as a foreign language at the English Language Institute, University of Michigan. Classes at varying levels of ability as determined by placement tests of English grammar and listening comprehension were tested. Subject responses were written descriptions of visual displays. The basis for scoring was information transmitted.

Two findings from this work were of special interest, and stimulated further work. First, we found that given enough time all subjects were able to write descriptions which would allow naive readers to make perfect reproductions of the displays being described. It was only when time constraints were imposed that individual differences began to appear. The second finding was that information transmission scores were consistently uncorrelated with a teacher's ratings of his students' writing ability, or with what he called *overall language ability*.

These findings suggested several general conclusions:

(1) that communication measures might tap abilities not generally considered by linguistically oriented teachers,

(2) that such abilities may compensate for linguistic competence as it is narrowly defined,

(3) that our current implicit definition of language performance might be too narrow, and

(4) that the naive grader technique might be a satisfactory means for more comprehensive production testing.

During the past year the feasibility of naive audience techniques for a variety of communication situations was investigated. I've already reported that they were found workable. In addition, an experimental set of eighty-six four picture items was prepared for a communication task in which the speaker is aware of the perceptions available to his audience. The subject's specific task was to communicate to a visually remote audience which one of the four pictures was identical to a single picture shown to him by an examiner. Subjects were instructed to try to communicate with one sentence. The picture sets were of two general types. In the first type the subject was instructed, *describe the correct picture*. The set contained pictures of four different objects, or actions being performed, four different values of some attribute, or more frequently the four conjunctive possibilities of two objects and two *actions*, etc. (e.g. the boy or a girl holding a flower or two flowers). The second type was similar to the first except for the addition of a representation of a man looking at or pointing to what might be the contents of a set of pictures of the first type. These were accompanied by the instruction, *What is this man doing?*

After an explanation and examples, the subject was given five seconds to look at each set of pictures. He was shown a duplicate of one of the pictures and had to indicate the corresponding picture in his set. After pointing out the correct picture, he was given the appropriate instruction to begin his communication. Total time for each item before the subject was ready to begin speaking was seven seconds. In the initial trial all subjects were able to identify the picture designated by the examiner.

Before trying out the test with subjects who did not speak English natively, four natives were given the test and their utterances tape recorded. Three different keys were used. Two naive native speakers listened to the recordings of all speakers and marked independently which of the four pictures in each set they thought the subject was indicating. No more than two communication errors were found with any combination of speaker and listener.

Ten students of English as a foreign language volunteered to take the initial 86-item test. Utterances were recorded. The primary purpose of this administration was to cull items from sets which elicited uniformly *identical* responses from all subjects. Five test forms were used with two subjects getting each of the forms. A secondary purpose of this administration was to get preliminary scoring reliability estimates for the technique and to get some idea about the relation of scores on this test to other test scores. Four natives listened to all ten test tapes in four different orders. All subjects took achievement tests in English grammar, vocabulary and listening comprehension within two weeks after the experimental test. Seven of the subjects took a proficiency battery including a written composition which was graded independently by three readers.

Two major measures were made with the first administration: (1) number of items correct, that is, the naive audience was able to select the correct picture and (2) total response time. The two measures correlated - .773. Mean interjudge reliability for items correct was .872. The lowest interjudge correlation for the derived measure of rate of communication (correct/time) was .989.

Composition and achievement test scores were correlated with raw scores, total response time and communication rate scores. Coefficients were uniformly high, greater than .8 in all cases, with the highest coefficients between communication rate on one hand and total achievement scores or written composition on the other (.9+).

On one hand it is embarrassing to report findings based upon such a small collection of data and discouraging to find such high correlations between a test which is difficult to administer and tests which are much easier to give. On the other hand, if the high scoring reliability is not an artifact dependant upon a fortuitous selection of subjects, and if the high correlations between communication test scores and other measures is due in part to the fact that the subjects had spent approximately two months in an English speaking environment, then an oral communication test of this form *may* have some value. At least it seems to be a possibility worth investigating. Therefore, on the basis of responses in the experimental administration the conceptual contents of the original eighty-six items were recombined in forty-six items. Most of the items can be represented by 2 x 2 tables in which the column and row headings represent pairs of different words with the same case relations. In order for each cell to represent a different sentence, cells usually all have identical context of word(s) with case relations.

For example, the set representing the sentences

The boy is sitting on the table.
The girl is sitting on the table.
The boy is standing on the table.
The girl is standing on the table.

can be formally represented as:

	Agent boy	Agent girl
Vsit	LOC on table	LOC on table
Vstand	LOC on table	LOC on table

This formalism allows us to describe the contents of a test form and to evaluate the comprehensiveness of sampling with reference to the base of a case grammar.

Further test development work is awaiting completion of a small logic system which can be used with existing language laboratories in order that (among other things) they may be utilized for administering tests of this sort. This will eliminate the need for individual test administrations, and will automatically provide a record of subjects' response times. In addition, it will allow a reduction in grading time to a range of ten to twenty seconds per item. It seems likely, therefore, that we can have a laboratory-administered test of oral communication with a reliability of approximately .95 which can be scored in an average of about 10 minutes per subject. This is quite obviously a guess; a year from now we should know how accurate it is.

Before I relinquish the floor there is one more speaking test I would like to talk about. When I was studying a foreign language during my first year in university, a part of the end-of-term examination was an interview conducted by a professor unknown to the students. Interviews were scheduled at fifteen minute intervals. Being cunning, if not an apt language learner, I decided on the best plan for getting an acceptable interview score. Since the interviewer would certainly begin by asking me my name, I prepared a lengthy genealogy; since, to put me at ease, he might ask where I was from or what I was studying, I rehearsed a geographical description and history of my home town and a lengthy summary of my studies. I was careful, of course, to include a few questions to elicit *yes* or *no* answers from the professor on order to maintain the semblance of conversation.

This audience may be surprised to learn that the plan worked. Up to a point at least. That point was some five or ten minutes after the scheduled end of my examination. What I had not reckoned on was that because my name began with *U*, my interview would be last; there was no other student waiting at the door to relieve me before my ignorance was exposed. That day is fresh in my memory.

REFERENCES

Bolinger, Dwight: *The Theorist and the Language Teacher.* In: *Foreign Language Annals,* II, 1 (1968), 30-41.

BRIERE, Eugène J.: *Testing ESL among Navajo Children, in Problems in Foreign Language Testing.* In: J.A. UPSHUR and Julia FATA (Eds.): *Language Learning,* Special Issue No. 3 (1968), 11-21.

CARROLL, John B.: *Fundamental Considerations in Testing for English Language Proficiency of Foreign Students.* In: *Testing the English Proficiency of Foreign Students.* Washington, D.C., Center for Applied Linguistics, 1961, 30-40.

CHOMSKY, Noam: *Review of Skinner's Verbal Behaviour.* In: *Language* 35, 1 (1959), 26 - 58.

CHOMSKY, Noam: *Aspects of the Theory of Syntax.* Cambridge, MIT Press, 1965.

COOPER, Robert L. : *An Elaborated Language Testing Model,* in: *Problems in Foreign Language Testing.* In: J.A. UPSHUR and Julia FATA (Eds.): *Language Learning,* Special Issue No. 3 (1968), 57-72.

DAVIES, Alan: *Oral Testing in West Africa.* In: Alan DAVIES (Ed.): *Language Testing Symposium.* Oxford University Press, 1968.

FILLMORE, Charles J.: *The Case for Case.* In: BACH, Emmon and HARMS, Robert (Eds.): *Universals in Linguistic Theory.* New York, Holt, Rinehart & Winston, 1968, 1-88.

GRICE, H.P.: *Utterer's Meaning, Sentence Meaning and Word Meaning*. In: *Foundations of Language*, 4 (1968), 225-242.

JAKOBOVITS, Leon A.: *Implications of Recent Psycholinguistic Developments for the Teaching of A Second Language*. In: *Language Learning*, XVIII, 1 and 2, (1968a), 89-109.

JAKOBOVITS, Leon A.: *Dimensionality of Compound-Coordinate Bilingualism*, in: *Problems in Foreign Language Testing*. In: J.A. UPSHUR and Julia FATA:(Eds.): *Language Learning*, Special Issue No. 3 (1968b), 29-55.

JAKOBOVITS, Leon A.: *A Functional Approach to the Assessment of Language Skills*. Paper presented at Idyllwild, California, November 1968c, (Mimeo.).

KLEIN, Sheldon and SIMMONS, Robert F.: *Syntactic Dependence and the Computer Generation of Coherent Discourse*. Santa Monica Document TM-758, 1962.

KLEIN, Sheldon: *Experiments in Automatic Essay Writing*. Santa Monica, SDC, 1964 (Mimeo.).

LADO, Robert: *Language Testing*. London, Longmans, Green & Co. Ltd., 1961.

LINDSAY, Robert K.: *A Heuristic Parsing Procedure for A Language Learning Program*. Information Processing Report No. 12, Austin, The University of Texas, 1964.

MCNEILL, David: *Developmental Psycholinguistics*. In: SMITH, F. and MILLER, G.A. (Eds.): *The Genesis of Language: A Psychological Approach*. Cambridge, MIT Press, 1966, 15-84.

MILLER, G.A., PRIBRAM, Galanter and K.: *Plans and the Structure of Behavior*. New York, Holt, Rinehart and Winston Inc. , 1960.

MILLER, George A.: *Language and Psychology*. In: HENNEBERT, Eric H. (Ed.): *New Directions in the Study of Language*. Cambridge, The MIT Press, 1964, 89-107.

MORTON, John: *Considerations of Grammar and Computation in Language Behavior*. In: *Studies in Language Behavior, Progress Report VI*. Ann Arbor, Center for Research on Language and Language Behavior, 1968, 499-545.

NEWMARK, Gerald and SWEIGERT, Ray L. Jr.: *A Field Test of Three Approaches to the Teaching of Spanish in Elementary Schools*. Cooperative Research Project No. D-177, Final Report, Sacramento, California, State Dept. of Education, 1966.

OSGOOD, Charles E.: *On Understanding and Creating Sentences*. In: JAKOBOVITS, Leon A. and MIRON, Murray S.(Eds.): *Readings in the Psychology of Language*. Englewood Cliffs, Prentice Hall, Inc., 1967, 104-127.

OSSER, Harry: *A Study of Communicative Abilities of Disadvantaged Children*. MD. (1968).

PALMER, Harold: *Principles of Language Study*. London, Oxford University Press, 1921.

PALMER, Leslie A. and UPSHUR, John A.: *Problems in Validation of Pencil-and-Paper Tests of Oral Production*. Mimeo, 1960.

PERREN, George: *Testing Ability in English as A Second Language: 3 Spoken Language*. In: *English Language Teaching*, XXII, 1 (1967), 22-29.

PERREN, G.E.: *Testing Spoken Language: Some Unresolved Problems*. In: DAVIES, Alan (Ed.): *Language Testing Symposium*. London, Oxford University Press, 1968, 107-116.

PITCHER, Barbara and RA, Jung B. with HARRIS, David P. and PALMER, Leslie A.: *The Relation between Scores on the Test of English as A Foreign Language and Ratings of Actual Theme Writing*. Statistical Report SR-67-9, Princeton, ETS, 1967.

QUILLIAN, R.: *Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities*. In: *Behavioral Science*, 12, 1967, 410-430.

RAND, Earl: *A Short Test of Oral English Proficiency*. In: *Language Learning* XIII, 3 and 4 (1963), 203-210.

RAPHAEL, Bertram: *SIR; A Computer Program for Semantic Information Retrieval*. Doctoral Dissertation, MIT, 1964.

SKINNER, B.F. *Verbal Behavior*. New York, Appleton-Century-Crofts, Inc., 1957.

SPOLSKY, Bernard: *What Does It Mean to Know A Language, or How DO You Get Someone to Perform His Competence?* Mimeo., 1968.

UPSHUR, John A.: *English Word Order Selection and Arrangement*. Unpublished, 1962.

VALETTE, Rebecca M.: *Modern Language Testing: A Handbook*. New York, Harcourt, Brace and World, Inc., 1967.

CRITERION-REFERENCED TESTING

Rebecca M. Valette

The new frontier in the measurement and evaluation lies in the direction of criterion-referenced testing. The aim of this paper is to describe criterion-referenced testing and to present a general blueprint for developing a battery of criterion-referenced tests in foreign languages. More specifically, we shall consider a model for the construction of criterion-referenced placement tests.

A. *Preliminary Questions*

Let me first take a few minutes, however, to define some key terms.

What is a criterion-referenced test? A criterion-referenced test consists of questions or items which are designed to measure whether or not a student has mastered a set of specific objectives. The most common example of the criterion-referenced test is the well planned classroom quiz (the *progress* test): the teacher, for example, assigns a list of forty French words with their English equivalents and the following day he dictates the English equivalents and expects his students to furnish the French words they have learned. If the teacher is using an audio-lingual approach, he might have his students memorize a dialog and then grade each student's performance on his recitation. In both of the above examples, the teacher has first established the *criterion* (the learning of a list of words or the memorization of a dialog) and then selected an appropriate means of testing whether the students have attained that objective. Frequently, however, the classroom test is a haphazard affair and not really a criterion-referenced test at all.

The distinguishing feature of the criterion-referenced mastery test lies in the nature of the test results and their interpretation. The student who receives a perfect score has met the criterion. For those students who do not receive a perfect score, it is possible to look at the items missed and diagnose areas of weakness. We will speak more about this diagnostic function of the criterion-referenced test a little later.

Another type of criterion-referenced test, in the more limited sense of the term, is a proficiency test which provides scores falling along a uni-dimensional scale called a GUTTMAN scale. A score of *3*, for example, on a 1-to-8 scale, may be interpreted to mean that the student has learned all the objectives 1 to 3 and nothing above three. Up to this time, the only subject in which this type of criterion-referenced test has been developed is in arithmetic: the child learns to add simple numbers before he learns to add longer numbers, etc.

What is a norm-referenced test? The standardized tests currently in use in the United States are all norm-referenced tests. In the area of second-language learning, this group includes the *Common Concepts Test*, the *MLA Cooperative Foreign Language Tests* in the four skills, the *PIMSLEUR Language Proficiency Tests*, the *College Entrance Examination Board Achievement Tests* in reading and listening comprehension, the *Graduate Record Examinations*, the *Graduate School Foreign Language Tests* and the *MLA Proficiency Tests for Teachers and Advanced Students*. All of these tests are accompanied with norm tables which allow the test user to compare the student's score (either raw score or converted score) with the scores of other students who have taken the examination. Generally these scores may be converted to percentile ranking, stanines or norm bands. The tests are so designed that a *perfect* score is indeed a rarity. In fact, if the norm-referenced test is administered to a group for which it has been designed, the resulting scores should tend to reveal the broader range and the bell-shaped configuration of the normal curve.

How does one distinguish between a criterion-referenced test and a norm-referenced test? The main difference between a criterion-referenced test and a norm-referenced test, that is, between a test which tells you how much a student does and does not know as opposed to a test which helps you rank a student with respect to one or several reference groups, lies in the manner in which these tests are constructed.

The building of a norm-referenced test usually entails the following steps: item specifications are developed, large numbers of items are written, these items are then pretested and analyzed in terms of difficulty level and discrimination ability, and the final form of the test is assembled so as to include items with the proper statistical characteristics. This final form of the test is sometimes validated against a criterion; for example, in the validation of the *MLA Proficiency Test for Teachers and Advanced Students*, the language tests of the battery were administered to adults in the foreign countries where those languages were spoken.

In constructing a criterion-referenced test, the first step is the establishment of a criterion; then specific items are written and validated against the criterion. The final form of the test contains those items which effectively measure the behaviors specified by the criterion. For example, let us imagine we are developing a criterion-referenced mastery test in French listening discrimination for American students. First we turn to a contrastive analysis of the sounds of French and English to establish a list of those phonemic and major phonetic differences which cause difficulty for American students. Then we write items to test these various problem areas. After the effectiveness of the items has been determined through pretests and item analysis, the final test is assembled so that each problem is tested at least once if not three or four times. Item difficulty and discrimination indices are not determining factors in the building of the final forms. On a listening discrimination test, for instance, it is hoped that all students will approach a perfect score after one year of instruction. The criterion-referenced test may furnish one raw score or, if it is designed for diagnostic purposes, several subscores. Once the criterion-referenced test is developed, it is of course possible to establish norm tables as for the norm-referenced test.

What is the need for criterion-referenced tests in foreign languages? In the United States we have many comparative measures of student achievement in foreign languages. These norm-referenced tests are effective for selection purposes, as in the case of the *College Entrance Examination Board Tests* and the *Graduate Record Examinations*. They are also useful to the researcher who needs measurement instruments to discover whether one teaching method or approach is more effective than another. But we have no criterion-referenced placement tests which can be used at the junior-high, high-school or college levels; we have no tests which can tell a teacher or the student himself what he does and what he does not know. Schools do now use some of the norm-referenced tests for placement purposes, but while this rough grouping is better than nothing, teachers are demanding a more usable battery of language classroom tests.

There are, moreover, two other specific areas where criterion-referenced proficiency measures are needed and where we have up to this time tried to make the best of an unsatisfactory situation by attempting *post facto* to equate norm-referenced tests with a criterion. One area is the evaluation of the reading ability of doctoral degree candidates and the other is ascertaining whether the language proficiency of a prospective teacher meets minimum standards.

The *Graduate School Foreign Language Tests* of the Educational Testing Service were conceived of as objective measures which would indicate whether or not an advanced-degree candidate could read research materials in a second language. The resulting tests, however, were norm-referenced tests whose scores were reported on a 200-800 standard scale. Most participating graduate schools set their minimal passing scores in the 400 to 500 range. A recent report by PHILIP HARVEY (*Foreign Language Annals*, Vol.II/1968/No.2, p. 165-173) describes a validation study in which the *GSFLT* was administered to fourth-semester college students: The fact that 88% of the French students and 92% of the German students in the sample scored above 500 leads one to wonder how meaningful the graduate school cut-off scores are. Clearly the lack of a valid and clearly described criterion against which to interpret the *GSFLT* scores means that the majority of the candidates who have been passing the test are, in fact, probably unable to use the second language as a tool in research.

The *MLA Proficiency Tests for Teachers and Advanced Students* also fail to provide the student with meaningful scores. The word *proficiency* in the title of the test is misleading for the battery is actually composed of seven high level achievement tests whose scores are primarily useful for ranking teachers and teacher candidates. If we review the history of this test battery, we note that the Modern Language Association did in fact establish a table of *Qualifications for Secondary School Teachers of Modern Foreign Languages* in 1954, and that the subtests of the MLA Proficiency Battery definitely correspond to the seven categories of the table: listening (aural comprehension), speaking, reading comprehension, writing, culture, applied linguistics (language analysis), and professional preparation. And yet, while the original intent of the proficiency battery was to provide schools and certification agencies with an instrument for determining whether teacher candidates had met the minimum standards set forth by the Modern Language Association, the actual tests were developed as norm-referenced tests. Consequently administrators have not known how to interpret the test results. Cut-off scores established by the states of New York and Pennsylvania only reflect certain percentile bands which have been established over several years of test administration. JOHN CARROLL (*FLA*, Vol.I/1967/No.2, p. 131-151) recently carried out a small-scale experiment in which he equated the proficiency ratings of the *United States Foreign Service Institute* with scores on the four language skills subtests of the MLA Proficiency Battery across the five languages involved: German, French, Italian, Spanish, and Russian. Although this preliminary equating study remains to be validated through a larger scale replication, it has become evident that not all tests cover the same range of proficiency. Perhaps the Modern Language Association will find it more advantageous to develop a criterion-referenced battery of proficiency tests rather than to attempt matching the existing tests to a criterion.

B. *Considerations in the Establishing of a Criterion*

The construction of a criterion-referenced test obviously presupposes the establishment of a criterion. The criterion itself is composed of one or several instructional objectives, however, we would do well first to consider the broad taxonomy of educational objectives and then draw up a table of specifications appropriate to foreign languages.

The Taxonomy of Educational Objectives

The BLOOM Taxonomy of Objectives, which has received rather wide-spread acceptance in the United States, lends itself to a classification of the objectives of second-language learning. The taxonomy is divided into three domains: the cognitive domain (which includes thinking and knowing), the psychomotor domain (which includes neuromuscular activities such as movement and speech) and the affective (which includes attitudes and appreciation). Table One presents a brief description of the categories in the three domains: the categories in the cognitive and the affective domain are borrowed from BLOOM and KRATHWOHL, whereas the categories in the psychomotor domain are my own (since no handbook on the latter domain has yet been published).

Learning a second language involves both the cognitive and the psychomotor domains, whereas other academic subjects such as history, literature, mathematics and science involve primarily the former, and the arts, such as ballet, sculpture and drama, involve primarily the latter. In the cognitive domain the student learns vocabulary items, sentence patterns, grammatical endings and sound-symbol associations (Category 1.0, *Knowledge*), he learns to understand the surface meaning of sentences and longer speech samples (Cate-

Table 1

TAXONOMY OF EDUCATIONAL OBJECTIVES

Cognitive*		Psychomotor		Affective*	
1.0	Knowledge	1.0	Perception	1.0	Receiving
2.0	Comprehension	2.0	Conscious Production	2.0	Responding
3.0	Application	3.0	Internalization	3.0	Valuing
4.0	Analysis	4.0	Interpretation	4.0	Organization of Value System
5.0	Synthesis	5.0	Creation	5.0	Characterization
6.0	Evaluation				

* taken from KRATHWOHL-BLOOM-MASIA

gory 2.0, *Comprehension*), he learns to express himself in the second language so as to convey direct meaning (Category 3.0, *Application*), he learns to understand the nuances and implications of different styles of language (Category 4.0, *Analysis*), he learns to use the second language to convey specific implications (Category 5.0, *Synthesis*), and finally he may reach the point where he can discern the stylistic merits and inconsistencies of selections in the second language (Category 6.0, *Evaluation*). In the psychomotor domain, we distinguish among written language, spoken language, and kinesics (or body language). If the second language uses the same writing system as the student's own language, he will only have to learn some new accents or marks of puntuation; the American students learning Russian, Arabic or Chinese must also learn how to write a new set of symbols, and this learning falls into the psychomotor domain. First the student learns to perceive the distinctive features of the new writing system, sound system or system of gestures (Category 1.0, *Perception*); then he begins to imitate and copy the new graphic symbols, sound patterns or movements, to perform them in the absence of a model, to manipulate them and, finally, to produce them, conciously, at will (Category 2.0, *Controlled Production*); with practice these new patterns become habitual (Category 3.0, *Internalization*). In the arts, for example modern ballet, the student may proceed to the next level where he can impart his own touch to a dance (Category 4.0, *Interpretation*) and then do his own choreography (Category 5.0, *Creation*). This distinction between the cognitive and the psychomotor domains parallels CARROLL's dichotomy between the cognitive-code-learning approach to foreign languages which emphasizes learning of vocabulary and the application of rules (either via a *grammar-translation* method, a *reading* method or a *direct* method) and the audio-lingual approach which stresses the formation of correct speech habits through dialog memorization, pattern drill and analogy. However, just as language teachers are realizing that effective second-language learning must combine intellectual understanding plus intensive practice to develop the activ skills, so do the architects of the taxonomy admit that in the actual learning situation there is a constant interplay among the three domains and the categories in those domains. None the less, the taxonomy categories are useful to the test maker in that they allow him to classify test items according to the kinds of behaviors they measure.

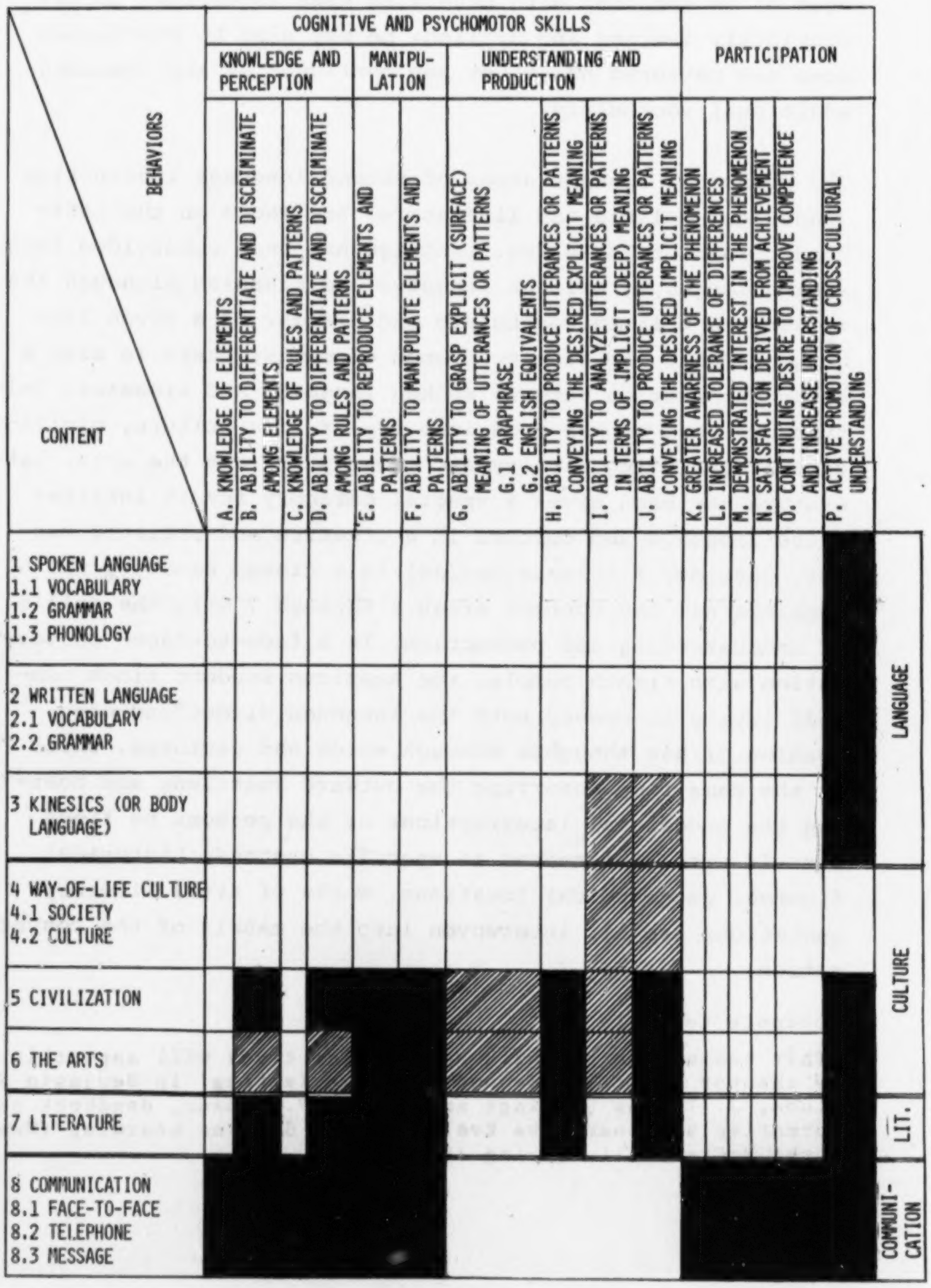
A Table of Specifications

The table of specifications permits a two-dimensional presentation of the interrelationship between the objectives and the content areas of second-language learning. Certain cells on the table have been designated *empty cells* or *improbable cells* to indicate that they are inappropriate in the usual second-language learning situation.

Objectives A through P are described across the top of the table. Objectives K through P refer to the affective domain; we shall not have the time today to consider this area. Objectives A through J concern the cognitive and psychomotor domains; A, C, G and I are specifically cognitive behaviors, B, D, E and F are primarily psychomotor behaviors, and H and J definitely include both cognitive and psychomotor components. The Objectives generally increase in complexity from A through J; however, some audio-lingual methods begin by teaching objective E (and sometimes F) before developing knowledge and perception. In learning a second language, the

TABLE OF OBJECTIVES

Table 2



■ EMPTY CELLS

▨ IMPROBABLE CELLS

BEHAVIORS / CONTENT	COGNITIVE AND PSYCHOMOTOR SKILLS										PARTICIPATION						
	KNOWLEDGE AND PERCEPTION				MANIPULATION		UNDERSTANDING AND PRODUCTION										
	A. KNOWLEDGE OF ELEMENTS	B. ABILITY TO DIFFERENTIATE AND DISCRIMINATE AMONG ELEMENTS	C. KNOWLEDGE OF RULES AND PATTERNS	D. ABILITY TO DIFFERENTIATE AND DISCRIMINATE AMONG RULES AND PATTERNS	E. ABILITY TO REPRODUCE ELEMENTS AND PATTERNS	F. ABILITY TO MANIPULATE ELEMENTS AND PATTERNS	G. ABILITY TO GRASP EXPLICIT (SURFACE) MEANING OF UTTERANCES OR PATTERNS G.1 PARAPHRASE G.2 ENGLISH EQUIVALENTS	H. ABILITY TO PRODUCE UTTERANCES OR PATTERNS CONVEYING THE DESIRED EXPLICIT MEANING	I. ABILITY TO ANALYZE UTTERANCES OR PATTERNS IN TERMS OF IMPLICIT (DEEP) MEANING	J. ABILITY TO PRODUCE UTTERANCES OR PATTERNS CONVEYING THE DESIRED IMPLICIT MEANING	K. GREATER AWARENESS OF THE PHENOMENON	L. INCREASED TOLERANCE OF DIFFERENCES	M. DEMONSTRATED INTEREST IN THE PHENOMENON	N. SATISFACTION DERIVED FROM ACHIEVEMENT	O. CONTINUING DESIRE TO IMPROVE COMPETENCE AND INCREASE UNDERSTANDING	P. ACTIVE PROMOTION OF CROSS-CULTURAL UNDERSTANDING	
1 SPOKEN LANGUAGE 1.1 VOCABULARY 1.2 GRAMMAR 1.3 PHONOLOGY																■	LANGUAGE
2 WRITTEN LANGUAGE 2.1 VOCABULARY 2.2. GRAMMAR																■	
3 KINESICS (OR BODY LANGUAGE)									▨	▨						■	
4 WAY-OF-LIFE CULTURE 4.1 SOCIETY 4.2 CULTURE									▨	▨							CULTURE
5 CIVILIZATION		■		■	■	■	▨	■	▨	■						■	
6 THE ARTS		▨		▨	■	■	▨	■	▨	■						■	
7 LITERATURE		■		■	■	■		■		■						■	LIT.
8 COMMUNICATION 8.1 FACE-TO-FACE 8.2 TELEPHONE 8.3 MESSAGE	■	■	■	■	■	■					■	■	■	■	■	■	COMMUNICATION

student himself finds himself functioning on various levels: when he is speaking with others he uses vocabulary he has previously learned and drilled; he may also be practicing some new patterns he is not yet confident of and learning additional vocabulary.

The principal content areas of second-language instruction (language, culture and literature) are shown on the left-hand column of Table Two. Language has been subdivided into spoken language, written language and kinesics although the spoken and written vocabulary and grammar of a given language usually reveal many common elements; there is also a close interaction between spoken language and kinesics. Culture has been broken down into *way-of-life* culture, civilization (including history and geography), and the arts. Literature has been given a special category for it interrelates language and culture in a creative and artistic manner. Category 8 (*Communication*) is a global category which combines all the content areas 1 through 7 with the skills of understanding and production. In a face-to-face conversation with French people, the American student finds himself trying to convey both the intended direct and deep meaning of his thoughts through words and gestures, while at the same time observing the outward reactions and hearing the occasional interruptions of the persons he finds himself with. References to specific customs, historical figures, geographical locations, works of arts, literary quotations are all interwoven into the matrix of the conversation.

Footnote to Table 2:

"This table together with sample item types will appear in my chapter entitled 'Second-Language Testing' in Benjamin S. Bloom, J. Thomas Hastings and George F. Madaus, Handbook on Formative and Summative Evaluation of Student Learning (New York: McGraw-Hill, spring 1970)."

Writing Specific Objectives

The teacher or test maker who plans to prepare a criterion-referenced test turns to the Table of Specifications as he begins to write out his instructional objectives. ROBERT MAGER has published a succinct handbook which describes the technique of writing specific objectives: first the overall behavior act is identified; then the conditions under which the behavior is to occur are defined; finally the criterion of acceptable performance is stated. One of the shortcomings of second-language teachers has been our failure to state our objectives in behavioral terms.

Let me try to clarify what I mean by a specific objective by giving you first a poorly worded objective and then the same objective stated in behavioral terms. The *Statement of Qualifications for Secondary School Teachers of Modern Foreign Languages*, which I mentioned earlier, defines the minimal level of listening comprehension (aural understanding) as follows:

> *The ability to get the sense of what an educated native says when he is enunciating carefully and speaking simply on a general subject.*

The test maker who wants to determine whether a group of teacher candidates in French have met this minimum standard will require some additional clarifications. The behavior act is obviously listening comprehension (cell 1-G on Table Two). Some of the conditions are given; we know that the speaker is to enunciate carefully, yet we must also know whether he speaks a sentence only once or whether he repeats it. We are told that the speaker is to speak simply, but here the test maker will need more information. Does this mean sentences may not exceed a maximum length, such as 8 words? Does this mean that no relative or subordinate clauses are used? Does this mean that only the present tense occurs in the sentences or perhaps the present, the *passé composé* and the future? Furthermore, what constitutes a *general subject*? A list of topics would be necessary; a vocabulary listing would be even more useful. The student is expected to *get the sense* of what is being said: Does this mean he should be able to recognize English paraphrase of at least 60% of material he has heard? Or does it mean that he should be able to identify the general topic of each recorded conversation that he listens to (100% performance on simpler items)? Is he expected to show his comprehension through English, or by identifying a written response in the second language, or by picking out an appropriate picture, or by giving a spoken response?

Let us now rewrite the above objective in specific behavioral terms:

Objective:	Listening Comprehension (cell 1-G)
Expected Behavior:	Given orally a conversation of three to five exchanges, the student will be able to identify the situation and the main topic of conversation. He may respond in his native language.
Conditions:	Conversation will contain only vocabulary and structure from *Français Fondamental* 1^{er} *degré*. Sentences will be enunciated clearly and spoken at slow conversational speed. Each conversation will be heard only once and the student will have time to respond after each conversation.

Minimal Acceptable Performance: 90% of the questions must be answered correctly.

It is evident that the student upon reading the second version of the objective, knows precisely what minimal performance is expected of him and that the test maker knows precisely what type of test to construct.

C. *Developing a Criterion-Referenced Placement Test*

A placement test evaluates the minimum levels of proficiency at predetermined points in a given curriculum. At the university level, for example, placement tests are used to section incoming freshmen with varying high school backgrounds in foreign languages.

Designing the Criterion

The first step in preparing a placement battery is determining the minimum content for each level or cut-off point. This content may be extracted from the textbooks and materials used in the school or adapted from some outside source, such as *Français Fondamental* or a curriculum guide. Cut-off point *A* for a French placement test might, for example, include the following content:

Vocabulary: the first 600 words of *Français Fondamental*

Grammar: the present, future and *passé composé* for all regular and irregular verbs in the list;
plural of nouns;
the determiners: *le*, *ce*, *un*, *mon*, *de* and numbers 1-100;

the agreement and position of adjectives in the list;
personal subject and disjunctive pronouns: *le* and *lui*;
object pronouns;
expressions of time, weather, age;
greetings and leave-takings;
sentences in the indicative, negative, interrogative;
sentences in the imperative.

Sound System:	major phonemic and morphophonemic patterns (which occur in the above vocabulary and grammar)
Writing System:	patterns of major sound-symbol associations (which occur in the above vocabulary and grammar)

Selecting a Format

Let us assume that School S, in the interest of economy has decided to include only listening and reading items on the placement battery (the active skills of speaking and writing are to be developed in the classroom). The *A* Placement Test might contain the following parts:

Part 1: Vocabulary
The student reads a short sentence in French. Then he is asked to indicate which of three English equivalents of an underlined word or phrase in the sentence is appropriate.

Part 2: Grammar.
Multiple-choice items test the student's ability to select appropriate grammatical forms needed to complete a sentence.

Part 3: Sound System.
The student hears brief sentences at slow conversational speed and must indicate, for example, whether a given sentence is a question or statement, whether it contains a singular or plural subject, whether it is negative or affirmative, whether the verb is in the present, past, or future, etc.

Part 4: Writing System.
The student hears brief sentences and selects the corresponding written form from among three options.

(Note: Each item of the test contains an *I do not know* option.)

The passing score for each part of the test is 90%.

Using Test Results

The student who passes all four sections of the test is placed in the upper class, while the student who fails all four sections is placed in the lower class. If a student passes parts 2, 3, and 4, but is weak in vocabulary, he is placed with the upper class but given remedial work in vocabulary. Students who pass parts 1 and 2 but obtain lower scores in parts 3 and 4 are placed in a special section of the upper class which will stress oral work. Students who do not pass part 2 are automatically placed in a lower class unless they have passed parts 1, 3 and 4. Students who obtain a perfect or a near-perfect score on the entire test should be given the more difficult *B* Placement Test.

A school may have several placement tests (*A*, *B*, *C*, etc.) of increasing difficulty. The specific placement test a student is given will be determined by his previous language experience. If the test he first takes proves to be much too easy or much too difficult, he is given the next higher or the next lower test, respectively.

D. *Modifications of the Criterion-Referenced Placement Test*

The criterion-referenced placement test may be modified to become a diagnostic placement test. An integrated sequence of placement tests might also become the core of a proficiency testing program.

The Diagnostic Placement Test

The four subtests of the model placement tests we have decribed above already furnish part-scores which help the teacher to diagnose each student's capabilities and weaknesses. Further diagnostic information might be made readily available if the test were administered with a double answer sheet. The student marks his answer sheet in the conventional way, but his answers are also transferred to an underlay which the teacher can use for diagnostic purposes: at a glance he can tell for instance whether the student is weak in verb forms or adjective agreement. A computer could also be programmed to deliver the same sort of diagnosis.

The Proficiency Testing Program

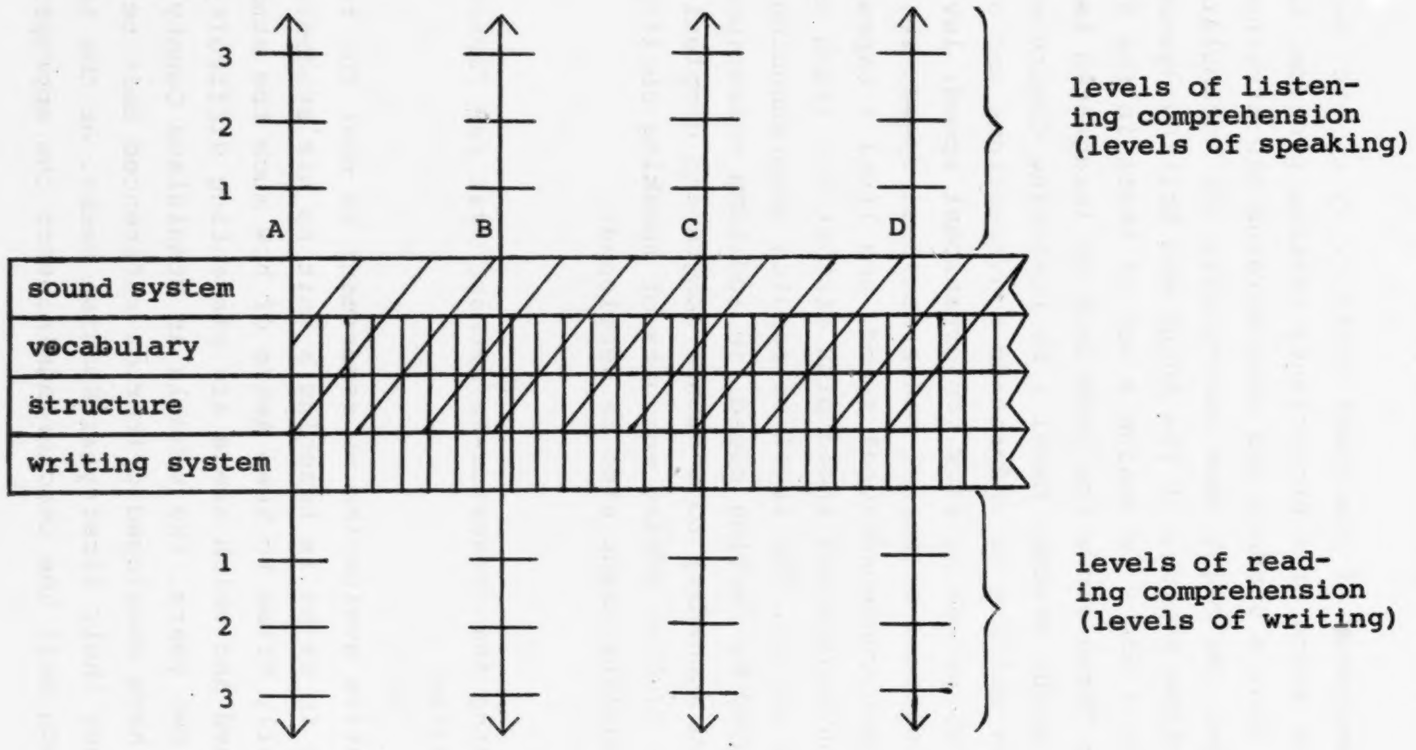
The sequence of placement tests *A*, *B*, *C*, etc. may be used as the score of a proficiency testing program. Let us assume that a student has demonstrated his attainment of the *B* level. We might then incorporate the vocabulary, grammar and those aspects of the Sound and Writing System which occur in *A* and *B* to design a set of tests in the four skills. Table Three shows the form such an integrated testing program might assume. Level 1 in Listening Comprehension would be the ability to understand conversations and other speech samples spoken at slow conversational speed; level 2 tapes are recorded at normal conversational speed, level 3 tapes at rapid conversational speed, and level 4 tapes are at rapid conversational speed plus distortions (such as static or white noise). The level of reading comprehension might be measured by reading speed; in addition selections would be graded according to sentence length and complexity of structure. Similar graded measures of speaking ability and writing ability might also be developed.

E. *Using the Criterion-Referenced Test for Formative Evaluation*

Formative evaluation or assessment is what the teacher engages in as he is teaching a unit to his students. He continually tries to keep aware of how much the students have learned and which areas are presenting difficulty. Over the past two years, the schools in Stanislaus County, California, have developed criterion-referenced unit tests to accompany their first-year Spanish texts. At the beginning of each unit the teacher administers the appropriate cri-

Table 3

MODEL FOR CRITERION-REFERENCED TEST BATTERY


3
2
1
A
B
C
D
sound system
vocabulary
structure
writing system
1
2
3
levels of listening comprehension (levels of speaking)
levels of reading comprehension (levels of writing)

terion-referenced test as a pretest. When the teacher feels that the students have mastered the unit, he administers the test again. However, unless 85% of the students score 90% or better on the test, the teacher cannot go on to the following unit. Rather, he continues with the current unit until his class demonstrates the minimum acceptable performance on the criterion-referenced test. This method of teaching to specific objectives has been demonstrated in California to increase the effectiveness of the teacher over one and a half times, and to benefit both the slower and the faster learners.

Conclusion

The innovative testing developments of the next decade will lie in the direction of criterion-referenced testing. In this paper we have described the difference between criterion-referenced and norm-referenced tests. We have presented a taxonomy of objectives: our own definition of the psychomotor domain was linked to the BLOOM-KRATHWOHL descriptions of the cognitive and affective domains. Then, we established a table of objectives in second-language learning which might form the basis for criterion-referenced tests. We also suggested the form a series of criterion-referenced placement tests might assume. Finally, we indicated how criterion-referenced tests can help teachers teach to specific objectives and thus improve their effectiveness. It is my hope that many of you in the audience will begin experimenting with criterion-referenced tests.

REFERENCES

Bloom, Benjamin et al., *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain.* New York: McKay 1956.

Krathwohl, David R., Benjamin Bloom and Bertram Masia, *Taxonomy of Educational Objectives, Handbook II: Affective Domain.* New York: McKay 1964.

Carroll, John B., Foreign Language Proficiency Levels Attained by Language Majors near Graduation from College. *Foreign Language Annals*, Vol.I/1967/No.2, p.131-151.

Harvey, Philip R., Minimal Passing Scores on the Graduate School Foreign Language Tests. *Foreign Language Annals*, Vol.II/1968/No.2, p.165-173.

Mager, Robert F., *Preparing Instructional Objectives.* Palo Alto, Cal.: Fearon Publishers 1962.

Smith, Melvin I., *Teaching to Specific Objectives.* Modesto, Cal.: Stanislaus County Schools Office, June 1968.

ORAL ENGLISH TESTING

Roy H. Kingsbury

Originally both JACK ALLSOP and PETER O'CONNELL were to have attended the conference, each, according to the original programme, to have led a discussion with demonstration. The original talks were advertised as:

JACK ALLSOP:

Some of the Problems of Oral Testing
(with examples from the AOE[1])

PETER O'CONNELL:

Towards a More Objective Assessment of Oral Production

Since neither JACK ALLSOP nor PETER O'CONNELL were able to attend the conference, I was asked to attend as a representative of the *Eurocentres*, and I undertook, at the same time, to lead a discussion on some of the problems of oral testing and cover as much of the two advertised talks as possible through a practical marking session of a candidate's tape from a previous *ARELS[2] Oral Examination*.

I repeated in brief the remarks made by JACK ALLSOP at the outset of his *Demonstration of the AOE* given at the *Göhrde Conference of Experts on Testing* in 1968 - a short history of the examination, the main reasons for its production and a description of the six sections which constitute the examination[3].

---

1) AOE = ARELS Oral Examination, which leads to the ARELS Diploma in Spoken English.
2) ARELS = Association of Recognized English Language Schools.
3) IRAL, Vol. VII/1969/No.2, p.136-138.

Before beginning the actual *marking session*, I gave a short survey of the conference held at *Folkestone* earlier this year which was devoted to the *ARELS Oral Examination* and its effect upon teaching. I stressed that one of the most valuable sessions at the conference had been the potential-marker training session, during which delegates had split up into groups to examine one candidate's tape from a former examination. I stated that, since such a practical session gave rise to discussion on all aspects of the testing of aural comprehension and oral production, I intended to follow the same type of procedure.

I then carefully explained the procedure I wanted to follow for each section: (a) *make the criteria clear*; (b) *mark a small number of items*; (c) *discuss any problems that arise through discrepancies*; (d) *complete marking*; (e) *discuss other problems*.

I pointed out, as JACK ALLSOP had done in a previous report, how we had aimed (a) *to make the examination easy to administer, namely by using a pre-recorded master-tape in the language laboratory*, and (b) *to make the examination as reliable as we could by employing objective testing techniques, clearly defined marking criteria, and by having each candidate's tape assessed by at least two independent examiners.*

I also outlined the six sections of the AOE:

(1) General Comprehension
(2) Situation Responses
(3) Reading
(4) Controlled Oral Composition
(5) Structures
(6) Free Oral Composition.

In *Section One*, the candidate hears a completely natural, unedited dialogue of about two minutes' duration and is then asked ten questions to test how well he has understood the dialogue. I stressed the fact that it is important that the questions test understanding and not merely the ability to recall facts.

In *Section Two*, the candidate is required to respond immediately, appropriately and with correct intonation to a series of twenty-five situations, such as:

Stimulus:	*I've just won fifty pounds in a competition!*
Response:	*Congratulations!* - or others, as appropriate;
or:	
Stimulus:	*You have been invited to a party and you wish to refuse without offending. What do you say?*
Response:	*That's very kind of you, but I'm afraid I have a prior engagement* - or others, as appropriate.

Section Three is primarily a reading-aloud section, in which the student is provided with scripts for the three subsections. In subsection one, he is required to take one of the parts in a telephone conversation, during the course of which he may be complaining, requesting, or expressing gratitude successively. His performance is assessed on predetermined bits of what he produces, with marks awarded for such elements as stress, intonation, flow, and tone. In subsection two he reads aloud (as if over the telephone) a letter which contains a wide variety of common abbreviations and symbols, including dates, numbers and letters of the alphabet. In subsection three, he reads a very short passage and is scored on a selected number of difficult words contained in it, such as *literature*, *particularly*, *produce* (noun) and *produce* (verb).

In *Section Four*, the candidate is provided with a ten or twelve-frame picture story which he studies and then retells. Instructions are quite detailed so that the candidate knows how much he is expected to give in the three minutes availabe. His performance is assessed according to such criteria as grammatical accuracy, appropriateness of vocabulary, fluency, etc.

In *Section Five*, a series of carefully structured questions are asked about the picture story which force the candidate to use particular structures in his response, such as compound tenses, passives, gerund constructions and indirect speech.

In *Section Six*, the candidate has previously prepared to talk persuasively for one and a half minutes, e.g. he tries to persuade his brother not to get married, or his parents to come to England for a holiday, or his friend to join the Conservative Party. The criterion is persuasiveness, and other factors are ignored except insofar as they reduce the effectiveness of the candidate's performance.

As I explained, a great deal of care goes into the administration of the examination, to reduce the candidate's nervousness at using electronic equipment, and in ensuring by lead-ins, examples and other devices that the candidate knows exactly what is required of him.

Although considerable thought and pretesting is involved in the preparation of the material for the examination, it is clear that the reliability of the examination depends largely upon:

(1) *the most precise and accurate definition of marking criteria;*
and
(2) *the training of examiners.*

The marking session then took up two whole morning sessions, with approximately 20 delegates listening to and assessing the tape of a former candidate from the November 1968 examination. I am very glad to report that delegates were extremely impressed by the examination and my impression was that such a practical session did much to highlight the real problems of aural/oral language testing.

Discussion centred firstly on the reasons for variations in assessments and secondly on the problems involved generally in testing the various factors of spoken English.

Although certain suggestions were made to increase the reliability of the instrument, the AOE was still recognised to be unique in its attempt to provide a complete examination at an advanced level in aural comprehension and oral production.

ERROR ANALYSIS

Alan Davies

1. In this paper I first consider three questions: What is a test (T), what is an error (E), and what is error analysis (EA). I then compare T and EA and argue that while there are important theoretical differences it is not clear what information EA can provide which T does not already give us.

2.1 What is T? Numerous definitions have been given. Here are three:

1. *An examination or test procedure, if it is to be efficient, must discriminate among the subjects examined or tested both reliably and validly.* (PILLINER, 1968 p. 23)

2. *A psychological or educational test is a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual.* (CARROLL, 1968 p. 46)

3. *... testing control of the (transfer learning) problems is testing control of the language.* (LADO, 1959 p. 24)

2.2 In the language field there are several kinds of tests; we shall consider only proficiency and diagnostic tests. They differ from one another in three ways, in purpose, in the group to be tested, and in the choice of language (sample might be a better term). The last two differences can be reduced to one, i.e. they depend on one another since we do not test someone who wishes to drive a car by putting him on a bicycle, nor do we test a child's knowledge of French by a task in Chinese. A test then is a set of tasks (usually related ones since a test is structured round the principle of homogeneity) given to a group of learners. Here we have two sampling problems, the sample of learners and the sample of tasks. The first is difficult but can be settled to some extent statistically. Given that we know the population we are interested in, then we try, so far as possible in making the test, to give it to all that population and in some cases we might be successful, e.g. all secondary school learners in Chinese in the U.K. - and even more easily if we restricted ourselves to, say, Edinburgh. But given that for most of the populations we are interested in we cannot get hold of everyone, then we must take a sample and it is accepted that if our sampling is random then we can make projections to what the results mean in terms of the population. The second sampling problem concerns the choice of tasks; if the first is difficult this is probably impossible. The problem here is not of being able to take a random sample, but of knowing what the population is in the first place. We do not know what are the discrete parts of language learning, nor do we know what their sequence is. What this means is that every test takes a set of bits of a language and assumes that they represent some part of what it is the native speaker can do. A proficiency test makes some sort of 'best guess' of what learners should be able to do at some particular point. A diagnostic test is apparently different in that it is made up of errors and here is the obvious link with error analysis. Diagnosis is a use of a test and we could therefore make use of a proficiency test for individual (or even group) diagnosis. However, it is more usual to think of a diagnostic test (even if we choose to think of it in terms of what the learner can do and re-term it a non-achievement test) as not being as structured as a proficiency test since a diagnostic test can be made up of a number of unrelated items which it has been shown learners of a particular type tend to get wrong, (and the statistical rationale behind diagnostic and other tests tends also to differ).

2.3 What does a test test? A test is not a questionnaire (though a questionnaire may eventually be the basis for a test) nor is it a survey; even less is it a set of exercises, or in any sense a teaching device. A test is a means of finding out something. The 'something' will depend on the use to which we may wish to put the test and we have already mentioned some of these. The 'means of finding out' also requires some explanation. Like any 'scientific' test or experiment instrument, it must, as PILLINER tells us above, be reliable and valid, (or, in LYONS' criterion for semantics, possess material and observational adequacy). Reliable means that any tester will obtain the same results; and there is the further implication here that the testees will produce the same results on successive occasions. The 'objective' item is essentially an answer to the demand of reliability. A reliable test may be very efficient but not test anything worth while. This is why a test must be valid too. Valid means that you know what you are testing and it is what you want to test; this can be assessed by correlation or content analysis techniques. There is the further implication of knowing what you want to test; viz.

that the ability under test conforms in some way to the theory of differential ability, i.e. that it can be described on a normal curve. Along with reliability and validity we shall consider this as a third characteristic of tests.

2.4 Let us now consider some examples of test items which may suggest to us just what it is a test tests:

1. Underline the word closest in meaning to the underlined word.

   Our tom cat has been missing for a week.

   wild
   drum
   name
   male

2. Here are three words. Are all the same, all different, or two the same?

   (1) bit (2) bit (3) beat

   All same.
   Two same: 1 and 2; 2 and 3; 1 and 3.
   All different.

3. Here are three sentences. Are all the same, all different, or two the same?

   It's 5 o'clock. - Read as (1) statement, (2) question, and (3) with double stress on 5.

   All same.
   Two same: 1 and 2; 2 and 3; 1 and 3.
   All different.

4. Listen to the following question and then choose the best answer.

What did John leave with you? (1) In the afternoon
(2) With his brother
(3) A large parcel
(4) Yes, he did.

5. Which of the following choices would a native speaker make?

I'm going to wash my hairs tonight.
the hair
my hair
hair

2.5 When we ask why items such as these are tested there are 3 answers:

1. They are critical points either in LADO's sense of this being points of interference for the learner or that they are caused by analogy. Whatever the cause they have been shown up as *Common Errors* (FRENCH, 1955).

2. They are easily testable. This is very important, perhaps the most important point. Saying the right thing in the right situation is very important to the learner but useless to the tester because it is untestable.

3. They are important to the learner at his particular stage, e.g. a physics lecture to advanced students in higher education. This is not a 'critical point'

in general but is very important to a particular group. On the whole most items will be determined on all 3 grounds. It is clear also that test items (my point 2 above) test reception and not production.

3.1 What is an error? In T, errors are what a diagnostic test tests. But then there is the further question of how these are established in the first place and this may be on the basis of a survey of the 'Common Errors' type or simply on the basis of a previous proficiency test. What is the difference between items which are 'error' full in a diagnostic test, and the ordinary run of items in a proficiency test. Clearly we are not going to put items in a diagnostic test which are impossible for anyone in our target population. Surely there is no difference except that the diagnostic test items are more *difficult* for the average learner. Now just as we do not fill a diagnostic test with impossible items, so we do not fill a proficiency test with trivially easy items such that all learners in our target population get everything right. This may be good for our egos as teachers but would sin against my third characteristic of what makes a test; such a performance would do, perhaps, for an exercise, but not for a test. We conclude, therefore, that what a proficiency test does is just what the diagnostic test does, it produces errors. There will be, of course, a differential as to how many errors any one individual learner produces but this just means that we have a test and not an exercise. It is also probable that in a diagnostic test we are more closely interested in a very homogeneous group and wish to distinguish individual errors and syndromes of errors. In a diagnostic test our interest is psycholinguistic. We are now in a position to modify our definition of error. An error is what average and below average learners regard as a right reponse to a test item.

3.2 It follows from what we said earlier about the demand of reliability that the errors produced will be consistent ones. A test is, after all, not a trick. It does not capitalise on slips of the pen or tongue (or their correlates in reception). The errors produced will therefore be systematic ones in the sense that they will tell us something about the competence of the learners. We can, therefore, make use of a set of test results to make profiles of what a group of learners possess as their competence, what is, in other words, their grammar and we can do the same for an individual learner.

4.1 We now come to our third question of what is error analysis. This has already been answered in CORDER's paper. To an error analyst an error represents a deviance from native speaker competence. CORDER of course points out that it is preferable not to speak of errors and to consider the whole production of a learner as potentially deviant, or, in his preferred term, idiosyncratic. I see his argument but it seems to me more reasonable to argue that if a learner produces a set of utterances which we must on all grounds, including contextual and situational, regard as non-deviant, we must ask ourselves how we judge them to be non-deviant and we must surely answer that it is in terms of their relationship to a grammar of native speaker competence. We consider errors therefore to be systematic deviances from native speaker competence and I should like to distinguish two senses of systematic: the first has to do with regularity; the second has to do with the system produced. The suggestion here, notably by CORDER (1967) is that every learning stage provides us with a system: this implies an infinite number of possible systems. I see this as an attractive hypothesis but it does seem to me as potentially vacuous since I can see no way in which we could ever describe an infinite number of systems or set up any kind of rules which would project them.

5.1 We are now in a position to make several distinctions between EA and T. *First* EA is individually oriented, it is more interested in psychological determinants than in group ones; (though presumably EA could be done for a group with assumed common characteristics); T is more concerned with group parameters and is therefore socially oriented. A social group is made up of individuals and individuals are of interest (unless they are poets) in terms of their language only if their behaviour can be generalised to some extent. We may state as the first difference between T and EA that T sets up some norm of group language behaviour and then compares an individual's performance (or that of a group) with that of the norm. EA in its investigation of individual language behaviour has no norm except assumed systematicness. But we must surely agree that it does also have the norm of native speaker proficiency in the target language with which it must implicitly compare the learner's behaviour.

5.2 *Second,* EA is a longitudinal method of study whereas T is a 'sample bite' method. Of course T can be used longitudinally but they still remain bites; this is also true of EA since no observation method is ever going to observe everything. However, the more observed the better the sample, on the whole.

5.3 *Third*, EA is open ended (and probably data oriented); it asks, 'What do you say?'; T is objective (and probably non-data oriented); it asks, 'Do you say X?'. However, we must distinguish between orientation and technique since EA could easily switch from 'What do you say?' to 'Do you say X?' and would probably wish to do so by confining the learner to certain fixed responses; at this point it would become very much like T in preparation; notice that I am *not* saying it would become a test because I wish to distinguish clearly between tests in preparation when they are necessarily speculative and tests in use (i.e. already constructed ones) which have now become 'scientific' instruments.

5.4 *Fourth*, EA is research oriented; it is not committed to a 'normal', 'average' way of learning because it is concerned with the individual; but this presumably means that there will eventually be useful generalisations, i.e. that there are patterns of learning. T on the other hand, assumes knowledge of what has to be learnt and also that there is a statable progression norm of learning. T therefore makes a 'best guess' as to learning stages. A good language proficiency test could be used diagnostically to show patterns of learning, and could even be used sequentially (i.e. at successive stages in the learning process).

5.5 *In conclusion*, what have we been saying: EA is clinical; *T is experimental*; EA is theoretical; T is practical; *EA is psycholinguistic*; T is educational. While T is experimental it is a procedure for experiment, an instrument, whereas EA is in itself a whole view. While T is clear about its methodology it is very uncertain how EA goes about its business. EA and T meet in two places: when EA is used as a technique itself (i.e. it has become 'Do you say X?') and then is on the verge of a test; and when T is itself a psycholinguistic operation.

5.6 *In a word then EA is concerned with grammatical description; T with placing individuals or groups within a grammar that has already been described.* From this point of view it is T that is data oriented and EA that is rule based. We may still wish to ask what rules EA can construct since they are potentially different for every learner. However, we have accepted that EA must be interested in eventual generalisation and therefore if EA could be made to produce statements as to learners' transitional competences in a useful way then the results would obviously be of great value to help in construction mere delicate or sophisticated, or simply just plain true T.

5.7 The basic difference between EA and T therefore is that they make different underlying assumptions. EA assumes that we do not know the grammar a learner has at any one time and proceeds to look for it (whether or not this can be done is another matter); T assumes that we know what grammar the learner is aiming at (as does EA of course) and, much more important, that it is possible to suggest a limited finite number of ways that will take him there. We have seen that EA and T agree about what has to be known (eventually); they disagree about the sequencing of arriving there. They further disagree as to how what has to be known, i.e. the language, may be broken down, or so it may appear. EA will define its units in terms of what the system imposes. While a T will test, say SV concord, or the negative transformation because they are there, EA will only incorporate them if they are relevant. However, this difference may be illusory because EA must start off by having the expectation that these features may be important; and T may reject them as part of a test for a particular group of learners (one nationality, say) for whom these features may not discriminate or be sufficiently accounted for by other parts of T.

5.8 Can T be used to provide us with grammars for transitional competences? T could be used to show sequential learning and give a sufficient number (infinite?) of transitional competences. But the only change that a test can show is *growth*. Once reliability has been established, further wrong responses to items previously got right must be treated as slips. (Incidentally, establishing reliability through Item Analysis is similar to discarding (in EA) slips of the tongue, pen etc.) T cannot handle as EA claims to (by contrastive analysis) changes which may include shrinkage. Consider a test of 60 items on the English Verb, each with 4 choices. Clearly the purpose of this T (as of EA) is to establish the learner's grammar (actually, more likely to establish its distance from a native speaker's). Now a test of only 9 items could produce over a million possible profiles or patterns of scores. It seems to me therefore unlikely that we shall ever arrive at either plausible grammars which are generalisable (60 items would produce enough individual combinations of different patterns for everyone in the world, present and past) or at acceptable ways of reconstructing our tests on each occasion to fit an individual's likely hypothesis.

5.9 But this is a pessimistic view. T cannot be as subtle as EA because it must be receptive while EA can be productive at least initially; I am not sure how long in technique it can remain productive. What this vast number of possible paths suggests surely is that T, given good items with plenty of overlap, can be used for exactly the same job as EA. Granted that T does make use of present assumptions as to grammar and sequencing; at the same time, since it sets up millions of conditions for any one state (progressively increased if we take states as a further multiplier), then T surely provides the range of possible grammars that EA claims and EA cannot claim more. We shall never, of course, analyse a million of these for T but then nor shall we for EA. What is pseudo procedure for one must be pseudo procedure for the other.

6.1 Let us push a little further this notion of what grammar T uses and in particular what is the nature of the distractors that are used in T. The purpose of Item Analysis, as we have said, is, in general, to make a test homogeneous; it maximises group patterns and avoids idiosyncratic ones; though there is dispute about this guillotine use of Item Analysis. Let us assume a test of, say, 20 items on Tense/Aspect, all of which discriminate highly; here are 4 of them:

1. John loves music whenever he hears it.
loved
is loving

2. Now that summer has come,
John has lived here for a whole year.
is living
has been living

3. Do you see that bright star a moment ago?
Did you see
Are you seeing

4. He had been studying in France for 3 years when war
has been studying broke out.
was studying

Let us consider an individual who responds half right, half wrong to all 20 items; what possible profiles can we suggest for him?

Here are three possible ones:

1. the choices under test (perfect, non-perfect, continuous, non-continuous) are in free variation;
2. perfect and non-perfect have coalesced as one or other;
3. continuous and non-continuous have coalesced as one or other; and so on.

6.2 T is inevitably limited by the range of distractors that it selects. Since it works to a known grammar and an assumed sequence it must use a limited range; what is more, its purpose of maximising group differences forces it to restrict and group the type of distractor it uses. T results, therefore, can show only that the learner has a partial, not a different grasp, on the assumption that T items all make use of the grammar of the target language. But if they don't (i.e. with distractors on the analogy of the learner's mother tongue or other languages he knows or other possible hypotheses such as teacher's eccentricity) then we might be able to show by T that the learner has a different and not just a partial grasp. The problem for T is *what* hypotheses and here is where we look to EA to provide some for us. But it remains difficult to see what can be the learner's hypotheses and how generalisable these may ever be if any of them are so idiosyncratic as to be concerned with, for example, teacher's eccentricity. Interference, yes; analogy, yes; other known languages, yes; and a good test will incorporate all these. And so we come back to partial knowledge, i.e. that the profiles we come up with are those of free variation or of a gap in the system. And we query whether EA is likely to succeed here simply because of the sheer practical difficulty. Our hope is, of course, that EA may come up with hypotheses which can be incorporated in new T.

REFERENCES

Carroll, J.B., *The Psychology of Language Testing.* In: *Language Testing Symposium*, ed. A. Davies, London: Oxford University Press 1968, p.46.

Corder, S.P., *The Significance of Learner's Errors.* IRAL Vol. 5/1967/No.4, p.161-170

French, F.G., *Common Errors in English.* London: Longmans 1955.

Lado, R., *Language Testing.* London: Longmans 1959.

Lyons, J. *Structural Semantics.* Publications of the Phililingual Society 1963.

Pilliner, A. *Subjective and Objective Testing.* In: *Language Testing Symposium*, ed. A. Davies, London: Oxford University Press 1968, p.19.

ZUR KONTROLLE UND BEWERTUNG MÜNDLICHER LEISTUNGEN IN DER ZERTIFIKATSPRÜFUNG

Heinrich Schrand

Die Diskussion über eine objektive Bewertung mündlicher Leistungen stand im Vordergrund der Arbeit, die die Gruppen Englisch und Französisch anläßlich der *Zweiten Internationalen Expertenkonferenz über Testfragen in der Erwachsenenbildung* leisteten. In den zum Volkshochschul-Zertifikat führenden Kursen soll ein gewisser Abschluß beim Erlernen der englischen bzw. französischen Sprache erreicht werden, das heißt, der Teilnehmer soll in diesen Lehrgängen zur Anwendung der fremden Sprache als Mittel der Kommunikation in seiner späteren Berufs- und Lebenspraxis befähigt werden. Deshalb werden besonders das verstehende Hören und Sprechen entwickelt. Beim Sprechen liegt das Schwergewicht auf der dialogischen Gesprächsfähigkeit.

Diese sprachpraktische Ausrichtung der Volkshochschul-Kurse zeigte sich vor allem in den Aussagen, die von den Konferenzteilnehmern zur Kontrolle und Bewertung der mündlichen Leistungen gemacht wurden. Man war der Auffassung, und das fand zum Beispiel in der von der englischen Gruppe erarbeiteten *Anleitung für den Prüfer* seinen Niederschlag, daß bei der Bewertung von sprachlichen Äußerungen einzig und allein Art und Grad der Störung der Kommunikation entscheidend sein sollten. Demnach sind künftig hinsichtlich der Verständlichkeit verschiedene Stufen zu unterscheiden:

Es können Fehler auftreten, die die Verständlichkeit in keiner Weise behindern.

Es können Fehler vorkommen, die die Verständlichkeit erschweren.

Einzelne Fehler stören die Verständlichkeit in hohem Grade; bestimmte Teile der Äußerung sind hingegen korrekt.

Die Verständlichkeit ist nicht mehr vorhanden.[1]

Der bisherige Mangel, keine einheitlichen Maßstäbe für eine differenzierte Fehlerbeurteilung zu haben, wurde nunmehr zum Teil beseitigt; denn die Konferenzteilnehmer hatten Erfolg bei ihren Bemühungen, Beurteilungskriterien, -prinzipien und -formen zu finden, mit deren Hilfe Fehlurteile im mündlichen Teil der Prüfung verringert werden sollen. Hinzu kommt, daß beim Englischzertifikat nunmehr auch nach Punkten bewertet wird, eine Regelung, die für das Französische von Anfang an bestand. Mit Hilfe dieses neuen Systems, das vom Positiven her an die Beurteilung der Leistungen herangeht, wurde die Voraussetzung dafür geschaffen, die Fehler zu differenzieren.

Im Mittelpunkt der Prüfung stehen die in der Broschüre *Das VHS-Zertifikat für Englisch*[2] festgelegten Zielfähigkeiten, die auch mündlich kontrolliert werden. Nach Meinung der Konferenzteilnehmer steht diese mündliche Prüfung im Hinblick auf die immer wieder betonte Neuorientierung des Erwachsenensprachunterrichts zumindest gleichberechtigt neben der schriftlichen. Das Prüfungsgespräch umfaßt die folgenden Teile:

---

1) Vgl. Helmut Süßmann: Einige Gedanken zur differenzierten Bewertung im Russischunterricht. In: Fremdsprachenunterricht. Berlin, Mai 1964. S. 253-254.

2) Robert Nowacek: Das VHS-Zertifikat für Englisch. Frankfurt 1968.

TEIL 1: *Lautrichtiges, ausdrucksvolles Lesen eines Textes*

Für die Vergabe der Gutpunkte ist die Gesamtleistung im Vorlesen eines Textes, besonders eines Dialogstückes, unter Berücksichtigung wichtiger Faktoren wie Zusammenhang, Bindung, Pausierung, Intonation und Betonung ausschlaggebend.
Erreichbare Punktezahl: 5.

Teil 2: *Aussprache von Testwörtern*

Der Prüfer stellt Fragen zum Inhalt eines Lesetextes, der Kandidat wählt aus zwei Antwortmöglichkeiten die jeweils richtige aus und liest sie vor. Jede Positivleistung in bezug auf die Aussprache des "kritischen Lauts" und die Antwortauswahl wird mit zwei Gutpunkten bewertet. Dabei ist der Prüfer auf die Beurteilung bestimmter Aussprachefaktoren, zumeist bedeutungsunterscheidender Laute (Phoneme), festgelegt, das heißt, er bewertet nur die Erscheinungen, die auf einem Bogen verzeichnet sind. Jede Positivleistung wird durch Abhaken festgehalten.
Erreichbare Punktezahl: 20.

TEIL 3: *Fragen zum Textverständnis*

In diesem Untertest wird die Fähigkeit des informatorischen Lesens sowie die des themagebundenen Sprechens überprüft. Es werden jeweils zwei Punkte für eine inhaltlich richtige Antwort vergeben, wobei wiederum ein Punkt abzuziehen ist, wenn die Antwort des Prüflings grammatisch unrichtig ist.
Erreichbare Punktezahl: 10.

TEIL 4: *Allgemeine Fragen*

Dies ist ein Gespräch in der Fremdsprache zur Beantwortung von Fragen zu alltäglichen Situationen, aus dem Lebenskreis des Prüflings, oder im Anschluß an Bilder bzw. Bilderserien.

In diesem Testteil werden für jede sinnvolle Antwort zwei Gutpunkte vergeben. Dabei ist jeweils ein Punkt abzuziehen, wenn die Antwort des Prüflings grammatisch unrichtig ist.
Erreichbare Punktezahl: 10.

TEIL 5: *Grammatikfragen*

Dieser Teil ist eine Art Patternbildungs-Test; grammatische Regeln werden nicht abgefragt. Für jede richtige Lösung wird ein Punkt vergeben, also kein Punkt für eine grammatisch unrichtige Antwort.
Erreichbare Punktezahl: 5.

TEIL 6: *Gelenkte Fragebildung*

In diesem Teil wird initiatives Sprechen geprüft. Es handelt sich um Gespräche, in deren Mittelpunkt Anfragen, Auskünfte und Wünsche aller Art stehen. Es wird jeweils ein Punkt für die grammatisch richtige Fragestellung vergeben; für eine grammatisch unrichtige Fragestellung erhält der Prüfling keinen Punkt.
Erreichbare Punktezahl: 5.

Schließlich können für die Teile 4-6 nochmals insgesamt bis zu fünf Gutpunkte vergeben werden, und zwar für natürliches und sprachgerechtes Verhalten in bezug auf Reaktion, Geläufigkeit, Verständlichkeit und Vollständigkeit.

Die vorstehend dargestellte neue Prüfungsform zeigt das Bemühen der Testfachleute, subjektive Interpretationsmöglichkeiten einzuengen. Es ist ja bekannt, daß schriftlich festgehaltene Lernresultate verhältnismäßig einfach zu messen sind, also auch objektiv zu bewerten. Problematisch ist dagegen die Beurteilung von Sprechleistungen, weil sich der Prüfer ganz und gar auf das nur einmal Vorgetragene konzentrieren muß. Die Prüfungsbeauftragten des Landesverbände sollen deshalb ab 1970 versuchen, der Lösung dieses Problems durch Einsatz eines Tonbandgeräts näherzukommen.

PICTURE COMPOSITION AND ORAL-RESPONSE TESTS

Terence Creed

The *Royal Society of Arts* was the originator of commercial and technological examinations in England, and still holds a foremost position in connection with the former. Instituted in 1856, they are now held throughout the British Isles and Commonwealth and cover all aspects of business studies - general, commercial, secretarial and modern foreign languages.

The Society first offered examinations in English for overseas students during the First World War, when in 1916 examinations were arranged for Belgian refugees. These continued until 1919. The total number of entries was 1043. In 1946 the Society was asked to provide examinations in English for Dutch students and these were provided from 1947 to 1949. They were held in The Hague under the control of the British Council.

It was in 1948 that the Society first considered offering examinations in English for Foreigners that would be essentially a practical test in the use of the language and available for all interested students. These were introduced at all three stages, Elementary, Intermediate, Advanced, in 1949. The syllabuses at the elementary and intermediate levels were concerned only with language work; set books were included in the syllabuses for the advanced examination but were discontinued after 1963.

When these examinations were first introduced it was intended that the examinations should be held in the United

Kingdom only, but in 1951 they were held also in Germany under the control of officials of the Control Commission. Since then there has been a steady increase in the number of entries coming from overseas centres under the control of officials of the British Council or of the British Embassy.

There has also been a parallel increase in the number of candidates taking these examinations in the United Kingdom. In 1949 there were 389 candidates from all centres but by 1967 the number had increased to 9780 and to meet the increasing demand the examinations at each stage are now offered four times a year.

From time to time the syllabuses have been reviewed to ensure that they are in line with the best current practice and in 1967 it was decided that a comprehensive revision should be undertaken. A circular letter was sent to a large number of centres (both in the United Kingdom and overseas) inviting comments and criticisms and many interesting and useful replies were received. A committee was then set up to make recommendations on new syllabuses. The members of the committee were all closely concerned with the teaching of the subject in a wide variety of establishments and included nominees of the *Department of Education and Science*, the *Association of Recognized English Language Schools*, the *Association of Teachers in Technical Institutions* and the *British Council*.

The great majority of students taking these examinations are from Europe and the Near East and stay in the United Kingdom for a limited period only - the average is about six months. Most are learning English for practical purposes in connection with employment. The number of candidates sitting these examinations in the United Kingdom form at least 3/4 of the total number of candidates. Centres with the greatest number of entries are found in London and the South East and all nationalities are represented.

The circular referred to above and sent to a large number of centres included questions, which are relevant to a discussion of written and oral composition:

E.G.

1. Would it be better to replace the composition by short pieces of continuous writing for which some guidance would be given (e.g. a letter or a description)?
2. Should the oral test be regarded as the most important part of the examination, carrying say 60% of the total marks? Some 65 out of 133 centres replied to this circular and a majority favoured a change from composition-on-a-theme to guided composition.

A majority of centres was opposed to weighting marks in oral-response test so heavily, i.e. at 60% of the total marks and suggested that there should be different weightings for different stages. This suggestion has not been adopted as yet largely because of uncertainty about 1) the importance of speaking in the acquisition of reading and writing skills and 2) the comparative importance of speaking at the three stages of a student control course.

The Revising Committee decided on the following description of the test to replace composition or writing on a theme:

The writing of continuous prose which may include

1) narrative or dialogue based on a given outline

2) giving an account of an everyday situation based on material supplied (This may be in the form of a picture, chart, etc.)

3) letter writing in a given situation;

and on the following description of its oral test:

> *The oral test will consist mainly of directed questioning of the candidates. The examiner will speak at normal conversational speed and may require the candidate to answer questions on some visual material provided.*

In the first series of the revised examination, the visual material was a picture-sequence used at Stage 1 to test both written and spoken composition. What I think is most relevant in our limited experience of these examinations, is the response to this kind of stimulus, the difficulties in administering the tests and the ways in which we might improve upon the stimuli used.

The factor common to stimulating both these production skills, speaking and writing, at Stage 1, was a non-verbal pictorial situation. The reason for this choice is the belief that pictures can give maximum of guidance and by providing the content about which the candidate should speak or write, so allow him to concentrate on expression. Although it might appear that this assumed division between form and content is not linguistically tenable, since thought gives rise to language and language to thought, nevertheless, it is a common experience to find students who are linguistically competent but have little to say, i.e. find it difficult to sustain unilateral writing or talking.

A considerable advantage in using pictures as stimuli to language responses under test conditions, is that the candidate can be given information without any, or with a minimum of words, so that he will need to find his answers from his own stock of language. This cannot be said of the story for aural reproduction, *Nacherzählung* or the summary-used-as-a-basis-for-composition, where words are used to produce words, often, unfortunately, the same words. An alternative is the writing-or-speaking about a given theme, but this involves advanced language skills and the validity of this stimulus is complicated by factors thought to be non-linguistic, e.g. imagination, experience and intelligence - this stimulus was rejected as producing responses that could be least objectively assessed.

The use of pictures, however, brings into the test situation a number of variables which may vitiate the validity and effectiveness of a test. The psychology of perception has demonstrated that graphic representations can mean different things not only to different nationalities, but also to different groups and individuals within a common cultural environment. For example, the absence of the conventions of perspective-in-drawing among primitive peoples has made inappropriate some of the learning materials produced by compilers from more sophisticated environments and we are often surprised by the variety of interpretations provided in our own countries by pupils and students who are learning through pictures. These difficulties in interpretation are real enough to have caused some examining bodies to discontinue the use of picture-sequences in tests of sustained writing. A precaution would be to pre-test pictures and this would seem to be the best way to avoid ambiguity, but is complicated by a number of facts. Firstly, pictures are specific, they are not easily resolved into patterns or types and the actual pictures to be used for testing would need to be pre-tested and therefore there is the danger that candidates may become forewarned. A bank of pictures might be built up so that the chance of a test-picture having been exposed to a candidate beforehand, would be reduced - in a way similar to that in which structural items are presented in bulk in multiple-choice tests, where a hundred or so may be worked at one time. After item analysis a number of the items may be included in the final test without much chance of students remembering these from pre-tests. Pictures, however, are not so easily presented meaningfully in such numbers and picture-sequences are more easily remembered by students than strings of words.

It may be useful, therefore, to consider some of the more obvious facts about pictures which have emerged from the *Royal Society's* revised examinations. Reference will be made to the picture sequences which were used in both the written and spoken composition. The sequences vary in quality of draughtsmanship from the primitive, sequence A, to the professional cartoon, sequence E. A sequence rather than an individual picture is presented in each, and only one exceeds four frames. Four of the sequences are humurous. The choice of pictures at Stage 1 has been determined by the following:

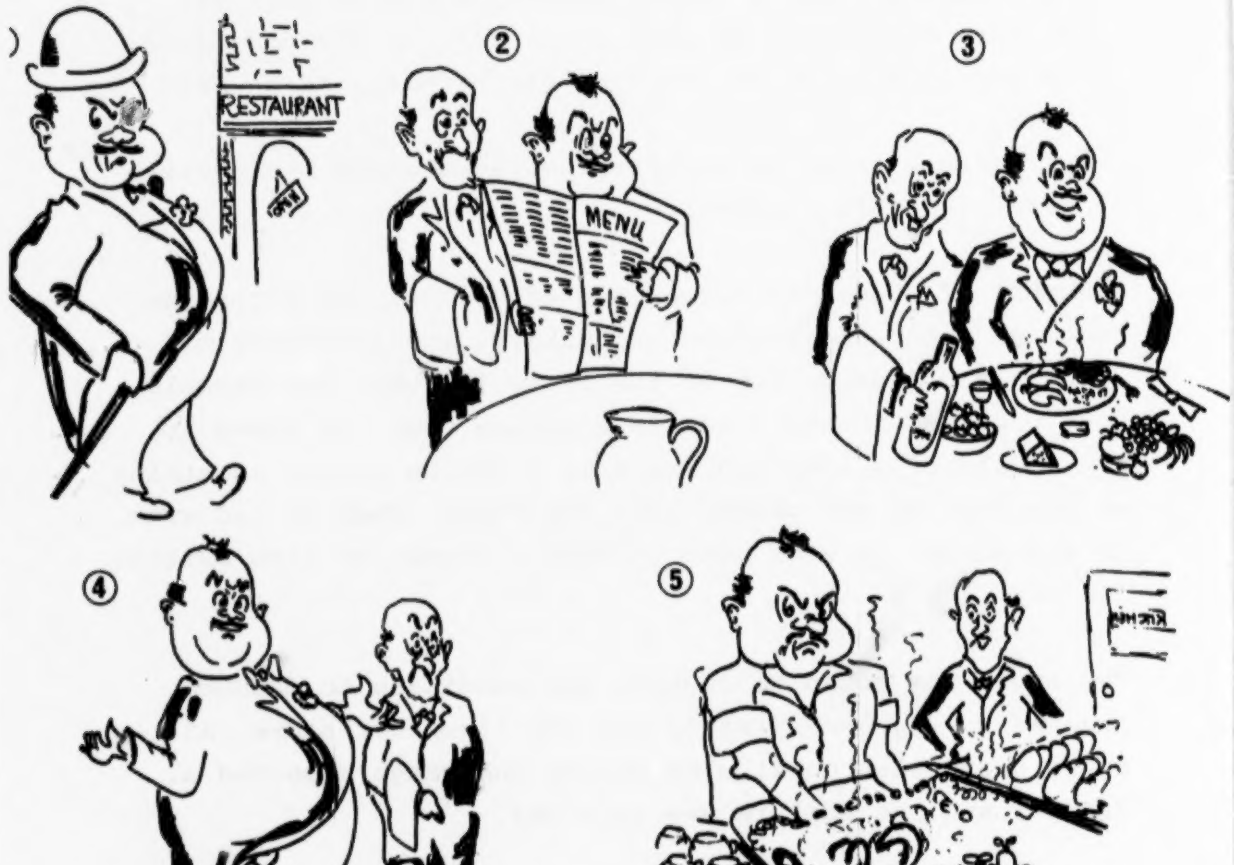
> *Simplicity and obviousness of the total situation;*
> *Humour is easy to illustrate briefly;*
> *The comic strip is an accepted convention with which*

candidates would be familiar;
Picture sequences tend to put an emphasis on action and verbs are among the heavy-duty words in English;
Picture sequences are easy to recall, if the candidate is required to answer without the stimulus being present;
A sequence tends to be less ambiguous since one picture tends to limit another.

A number of sequences were rejected because the situation required a too specialized vocabulary, the frequency of which was inappropriate to the Stage 1 level. For example, four pictures showed a man in a rowing boat. He rowed to the middle of a lake and began to fish. He caught something so big that he was pulled into the water. When he had swum to the shore, he went home to find a dinner of fish waiting for him.

The story was found to require the candidate to produce *row/oar/rowing boat/fishing rod and line/bait/shore*. All these items were considered beyond the range expected at Stage 1 and the picture was rejected.

It will be useful here to consider a significant factor already referred to in the choice of pictures as stimuli namely obviousness of the total situation. Only 49% of students tested on *Royal Society of Arts* picture I sequence provided the interpretation agreed by the examiners and intended by the cartoonist, namely a customer is forced to wash up in order to earn the meal he has not been able to pay for. The other interpretations are so varied that as a whole, they tell us nothing except that they confirm our belief in the variety of human perception - for example one


RESTAURANT
②
MENU
③
④
BILL
⑤

candidate spoke at length, *he had a marvellous meal and enjoyed it very much. Come on then, he said, let's wash up now.* The confused examiner asked, *Who is he? - The Manager,* answered the candidate.

This kind of confusion can be explained post-facto, by the inadequacy of the drawing. For example, is the man in picture 1 wearing tails, a bowler hat and carrying a walking stick? Is he standing or sitting in picture 2? Why hasn't the table been laid? Is that a handkerchief hanging out of his trouser pocket in picture 4? Has the waiter's face changed throughout the sequence, or is he maintaining a Jeeves-like imperturbability?

The pictures are further complicated by detail. For example, what is the significance of the button-hole in the customer's jacket and the disappearing handkerchief in the waiter's top-pocket? Why obscure the wine glass by nuts or whatever they are in picture 3, and is there any point in being fussy about how the word kitchen is to appear in picture 5 when chairs are missing throughout the cartoon?

Pictures for language teaching and testing at an elementary level need to present a simple situation in the clearest possible way.

A third factor influencing the obviousness of a pictorial situation is the degree of cultural-loading. Sequence II may not be understood by candidates who are not used to tax increases on tobacco and some candidates who interpreted sequence II assumed that either the convict had escaped to have a cigarette after which he dutifully returned to the prison, or the shop was closed and no amount of knocking

BACCONIST
CIGARETTES
NOW
5/7
FOR 20

in picture 4 appeared to persuade the shop-keeper to open up.

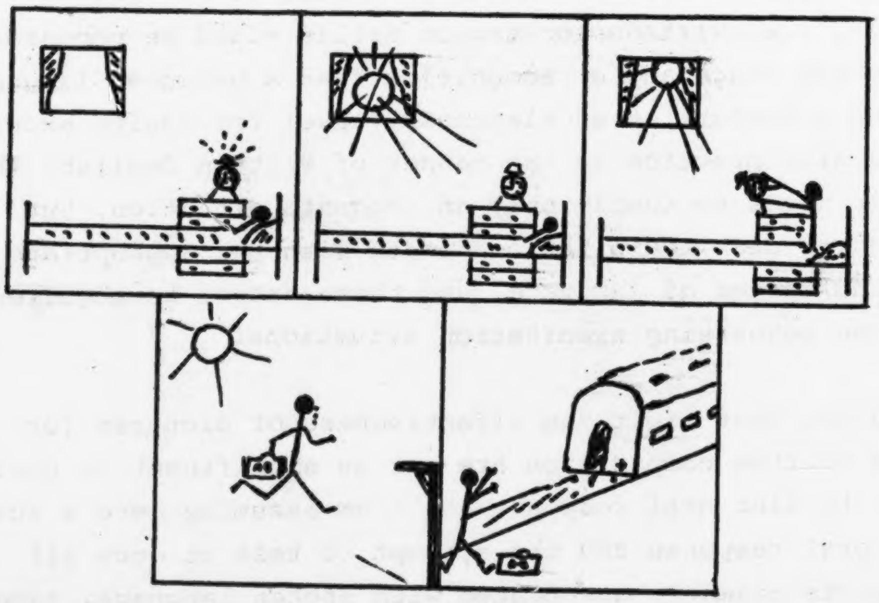
Similarly a U.K. examiner before briefing on sequence I wondered why the waiter should be helping his customer to the sauce in picture 3. Having missed the obscure wine glass, his environmentally conditioned eating-habits lead the examiner to believe that the bottle in picture 3 contained Worcester or some such sauce and sauce bottles are frequently found on tables in humbler restaurants. These cultural variables are less likely to arise in tests designed for Adults learning English as a Foreign Language and speaking not the same vernacular because it is likely they will have much in common, but it is as well to make sure that the situation is not exclusive to a professional or socio-economic group.

A general view among the examiners at Stages 1 and 2 (Elementary and Intermediate) was that the quality of presentation of the pictures (i.e. whether the pictures were professionally drawn or not) affected the attitude of the candidate to the tests. The accepted fact that language tests should have some face validity would seem also to apply to pictures when used as part of tests. Candidates reacted more quickly and more positively to a well-drawn picture. The view held by many teachers and students that pictures are useful only at the initial stages of learning and that they are at an advanced level an insult to a learner's mastery - this subjective view seems to find an echo in the candidate's reaction to the picture or picture-sequence, so obviously contrived. It is wise here to distinguish the picture for testing from that for learning; the latter can be supplemented by the teacher who can make the most ambiguous things clear by his questioning and commentary. In testing, at the elementary stage, however, most of the information will be presented pictorially and the oral examiner will not wish to waste much time in establishing a sympathetic attitude from the learner towards the stimulus. In written composition, the motivation of the candidate rests on the picture-sequence alone, in which case, it is all the more important that the candidate feels he is being taken seriously - and looked upon as an adult.

It has been implied in what I have said, that the advantage of pictures is that they present immediate information in a non-verbal form so that the candidate cannot borrow from the stimulus the language he requires for his answer. Nevertheless, the attempts of the majority of candidates at Stage 1 show in the written test a lack of organisation, a lack of an ordered and smoothly developed expression in words of the situation in the pictures. Most candidates describe the pictures in succession and very few internalise the information and reconstruct it in terms of cause and result which is the relationship most frequent in the elementary situations. This is obvious in the scripts where the candidate numbers the paragraphs each to correspond with a picture, or begins, *In the first picture* ... and continues *In the second picture* ... etc. The result of his arrangement is a good deal of co-ordinate clauses and compound sentences. Moreover, since the candidate observes rather than imaginatively participates in the situation, the verbs are nearly all in the present continuous and the word order tends to be monotonous, i.e. there are few inversions for emphasis, or introductory adverbial expressions.

Similarly such scripts contain a number of constructions which are more typical of spoken eye-witness commentary than written descriptions, e.g.

> *Somebody is in a room...*
> *We can see that he did not hear the alarm clock ringing before...*
> *In the last picture we can notice...*



On the other hand scripts at Stage 3, where aural recapitulation *(Nacherzählung)* is used as a test of sustained writing, show an orderly development more typical of written prose.

Perhaps the obvious answer to this problem of influencing the candidates to internalise before they produce, is to take away the picture sequence after they have studied it for say three to five minutes. Moreover, it may be wise to include some kind of set oral commentary on the pictures while candidates are studying them before beginning to write - a commentary which would concentrate on explaining the likely areas of ambiguity. In this way the variables in the stimulus are more likely to be controlled and a restrictioning in terms of prose conventions encouraged. These considerations throw up the question of what the features of prose at an elementary level are and how to train students to master them. It seems that a good deal of re-writing practice, i.e. written conversion drills would be necessary at an early stage and a recognition that a balanced language teaching procedure at an elementary level for adults should include also practice in the manner of written English. The examiner needs to supply only an adequate situation, but the teacher must equip the candidate with the appropriate stock and manner of language, and these cannot be acquired simply be rehearsing examination situations.

The factors that limit the effectiveness of pictures for testing written composition are not as significant in their use for testing oral response and I am assuming here a sustained oral response and the attempt to test at once all the aspects commonly associated with spoken language, namely the recognition and production of phonemes and prosodemes, fluency, and the mastery of the conventions or oral exchange in face to face communication.

Such an attempt may be considered unpracticable and pictures used to ascertain more discretely and objectively, e.g. adaquate discrimination between very specific items, e.g.

ship/sheep
washing/watching

but at some stage the candidate has to be assessed for his ability to converse and the interview or conversation situation will need to be stimulated. In the Society's exams at Stage 1, stimulating the candidate to speak is more successful than stimulating him to write. The presence of the oral examiner and his questioning add a third factor to the two in the written test, namely the picture stimulus and student response. Ambiguities and inefficient presentation can be compensated for by an examiner who is briefed to establish a sympathetic attitude from the student towards the test, and to guide the extent and kind of language of the student through graded questioning. This kind of questioning is a priceless heritage of language teachers and derives from GOUIN-PALMER and GUMEY. Hence the Society's examiners are drawn where possible from trained teachers of *English as a Foreign Language* who should be familiar with graded or staged questioning.

The first and most elementary kind of question is one that requires a *Yes* or *No*, the second requires a choice of alternatives with perhaps some modification of the structure given in the questions, e.g.

Did Mr. Jones pay his bill or wash up?
He washed up.

The third kind of question uses interrogative pronouns or adverbs and may provide little or no language which the candidate can use in his answer.

Why did Mr. Jones eat such a large meal?
Because he was hungry.

The first kind of question tests recognition, the second structural adaptation as well as recognition and the last, how much appropriate structure and lexicon the candidate is able to produce. The responses to the second and third questions are long enough to allow the examiner to sample phonemes and prosodemes. This sequence of questions also meets the need to help the candidate adjust gradually to the test situation and to explore with the examiner the pictures presented. Candidates are tested mostly in evening school-classrooms and conditions do not allow them to study the picture before they come into the test room. Hence the need for a minute or so of orientation before the examiner begins to make his assessment. Examiners are asked to refrain from obvious note-taking or marking since this tends to put the interviewee off and the practice of having the next candidate in the room while marking the outgoing candidate is discouraged, since the intervening silence appears to prejudice performance also. We have not yet experimented with assessing individual candidates within group conversation, but such an arrangement has much to commend it - it is less artificial, the stimuli are more varied and it is likely that the conversation will gather momentum quicker. Its disadvantages, however, are clear - personality factors will tend to intrude more, control of the stimuli is more difficult, two or three assessors would need to be present in the testing room.

The present oral examiners are also briefed on the need for making the time of the picture story clear to the candidate by cueing him in with, e.g.

This happened yesterday.
This has just happened.
What is happening?

If this is not done, most students use the present continuous. The examiners are also required to make sure the characters in the story have names so that they can be more easily distinguished. They are also told how to adjust the test to the immediate ability of the candidate. Since the test situation is designed to stimulate, the candidate who is immediately responsive may be asked to relate the story before much questioning. The not so responsive candidate may be put graded questions without prior narration. The least responsive may need to have the story related briefly be the examiner before he begins questioning.

The marking scheme for oral tests Stage 1, includes *comprehension, communication, intonation* and *stress, pronunciation* and *fluency*. You will not wish me to take up the discussions already familiar to you about such terms as *communication* and *fluency* and what they stand for, but it was agreed by the examiners that *communication* meant the appropriateness of both the student's responses and his questions to the stimuli (i.e. to the examiner's language and to the pictures). It was agreed that *fluency* meant ease of transition and a not too slow a speed - perhaps because of the inadequacy of the definition it received the least significance and the lowest weighting.

Examiners were advised to sample the candidate's spoken language for one aspect at a time and to begin this sampling once the candidate had reached a fairly even level of performance. Although minimal pairs are mentioned in the instructions to examiners these were not used since it was agreed that individual phonemes play a relatively small part in the total signalling systems of spoken language and that contexts where a minimal sound made a difference in meaning were very rare in everyday language. The lessons learned from the oral examinations were:

> *examiners need careful briefing including trial assessment of candidates on tape;*
>
> *they can cope with about twelve individual interviews in about 2 1/4 hours;*
>
> *they should be equipped with a tape-recorder inconspicuously positioned so that they can post-monitor the student's performance and their own questioning;*
>
> *they need to know the mistakes that students of a particular vernacular tend to make, so that they can direct their questions accordingly and judge how successfully the student has overcome his mothertongue habits.*

ZUR ENTWICKLUNG EINER HÖHEREN STUFE DES VOLKSHOCHSCHUL-SPRACHZERTIFIKATS FÜR DEN BEREICH WIRTSCHAFT UND INDUSTRIE

Robert Nowacek

Seit Jahren wurde in den verschiedensten Wirtschaftsgremien immer wieder darüber Klage geführt, daß die Fremdsprachennoten in den Zeugnissen der deutschen allgemeinbildenden Schulen kein verläßlicher Nachweis für den Besitz aktiver und praktisch verwertbarer Fremdsprachenkenntnisse seien. Es wurde wiederholt die Forderung nach einer bundeseinheitlichen Prüfung gestellt, die einen Grad an Fertigkeit im mündlichen und schriftlichen Gebrauch der Fremdsprache bestätigt, der eine sprachliche Behauptung in allen Lebenslagen möglich macht.

Da dieses Problem von der Kultusministerkonferenz nicht aufgenommen wurde, sah sich der Deutsche Volkshochschul-Verband veranlaßt, im Rahmen der Erwachsenenbildung für eine praktische Lösung zu sorgen. Das hatte den Vorteil, daß keine der traditionsbelasteten überkommenen Prüfungsformen des deutschen Schulwesens übernommen werden mußte, sondern daß man völlig neue Wege gehen konnte. Eine Grundvoraussetzung war, die Prüfungen so zu gestalten, daß ihr Bestehen - gleichgültig, an welchem Ort und nach welcher Vorbildung die Prüfung abgelegt würde - den gleichen Leistungsstand verbürgte. Es war unumgänglich, von den Erkenntnissen der Psycholinguistik auszugehen und die Prüfungen nach modernen Testverfahren auf den Grundsätzen der objektiven Leistungsfeststellung aufzubauen. Darüber hinaus sollten sie auf einheitlichen Prüfungsbestimmungen, Prüfungsaufgaben und Bewertungsmaßstäben beruhen. Als erster konkreter Schritt zur Verwirklichung des Projekts galt es, unter Mitarbeit bedeutender britischer Linguisten wie W. STANNARD ALLEN, A.S. HORNBY und MICHAEL WEST, im Ausschuß Grundlagenforschung das für die praktische Beherrschung nötige sprachliche Rüstzeug so klar wie möglich zu definieren und zu diesem Zweck verbindliche Aufstellungen über einen *Mindestwortschatz* und die wichtigsten grammatischen *Grundstrukturen* zu erarbeiten.

Nach einer erfolgreichen Erprobung in Bayern wurde im Jahre 1967 das Volkshochschul-Zertifikat für Englisch bundeseinheitlich eingeführt. Die Idee und die sorgfältige wissenschaftliche Untermauerung der Volkshochschul-Zertifikate durch den Internationalen Expertenausschuß fanden auch ausserhalb der Bundesrepublik starken Widerhall. Das *European Bureau of Adult Education*, die Spitzenorganisation der Europäischen Erwachsenenbildungs-Institutionen, identifizierte sich so weitgehend mit diesem Projekt, daß ihr Delegierter am 14. November 1967 bei einer Sitzung des *Council for Cultural Cooperation* in Straßburg die Forderung erhob, der Europarat möge sich dafür einsetzen, das Volkshochschul-Sprachzertifikat zu einem einheitlichen *Adult Education Certificate in Modern Languages* der europäischen Erwachsenenbildung auszubauen, um auf diese Weise jeder an der Übernahme der Sprachzertifikate interessierten nationalen Erwachsenenbildungsinstitution den Zugang zu öffnen.

Zur Verwirklichung dieses Vorhabens lud der Deutsche Volkshochschul-Verband in Verbindung mit dem Europarat Wissenschaftler aus 18 Nationen für Juni 1968 zur *First International Conference of Experts on Language Testing in Adult Education* ein. Auf Antrag dieses Gremiums wurde die Möglichkeit einer Übernahme der Volkshochschul-Sprachzertifikate nicht auf die Mitgliedsländer des Europarats beschränkt; durch die Erweiterung des *European Committee* auf ein *International Committee for the Adult Education Certificate in Modern Languages* wurde auch für außereuropäische Nationen die Einführung der Zertifikate grundsätzlich möglich gemacht. Mag dieser Beschluß zunächst nur Symbolcharakter gehabt haben, so erwies er sich bereits im darauffolgenden Jahr als eine entscheidende Voraussetzung für den positiven Verlauf der Verhandlungen mit amerikanischen Bundesverbänden über die Übernahme des zur Zeit noch in der Entwicklung befindlichen Volkshochschul-Zertifikats *Deutsch als Fremdsprache* (Deutsches Sprachzertifikat).

Inzwischen wurden in der Bundesrepublik bereits drei Prüfungen zu Volkshochschul-Zertifikaten für Englisch durchgeführt; außerdem fanden 1969 die ersten bundeseinheitlichen Prüfungen zum Volkshochschul-Zertifikat in Französisch statt. 1970 werden die Prüfungen in beiden Sprachen auch durch die Volkshochschul-Verbände in Österreich und in der Schweiz abgenommen werden. Hinzu kommt die Universität Clermont-Ferrand als Prüfungszentrale für Frankreich, die bereits 1969 die französischen Prüfungen abnahm.

In dem Bestreben, die weitere Entwicklung der Volkshochschul-Sprachzertifikate in größtmöglicher Praxisbezogenheit durchzuführen, waren der Deutsche Volkshochschul-Verband und seine Mitgliederinstitutionen darauf bedacht, die Kontakte zu Wirtschaft und Industrie zu vertiefen. Als besonders wertvoll erwies sich dabei der Gedankenaustausch mit dem *ERFA-Ring Wirtschaft*, dessen Mitglieder als repräsentative Vertreter der deutschen Wirtschaft gelten können. In diesem Kreise wurde dankbar anerkannt, daß die bestehenden Volkshochschul-Sprachzertifikate in mustergültiger Weise den Bedarf nach einem einheitlichen Nachweis solider, praktisch anwendbarer Grundkenntnisse einer Fremdsprache decken.

Vor diesem Gremium wurde auch die auf skandinavische Initiative zurückgehende Resolution der *Internationalen Expertenkonferenz* erörtert, eine zweite, wirtschaftsorientierte Stufe des Volkshochschul-Sprachzertifikats *(Higher Adult Education Certificate in Modern Languages)* zu entwickeln. Mit Nachdruck wurde auf die Dringlichkeit der Realisierung dieses Vorhabens vom Standpunkt der deutschen Industrie hingewiesen.

Auf Grund der Anträge der *Second International Conference of Experts on Language Testing in Adult Education* (Sonnenberg, April 1969) legt der Deutsche Volkshochschul-Verband folgende Planung für ein solches Zertifikat - zunächst für die englische Sprache gedacht - vor:

In Parallele zu den bestehenden Volkshochschul-Zertifikaten für Englisch und Französisch soll auch die zweite, wirtschaftsbezogene Stufe auf

a) einheitlichen Stoffplänen,
b) den Grundsätzen der objektiven Leistungsfeststellung,
c) einheitlichen Prüfungsbestimmungen, Prüfungsaufgaben und Bewertungsmaßstäben

basieren.

Selbstverständlich darf das angestrebte *Higher Certificate* in seinen Leistungs- und Stoffanforderungen den Rahmen der voll ausgebauten Volkshochschulen nicht sprengen, d.h. Inhaber der ersten Stufe des Zertifikats müssen das *Higher Certificate* in einem Arbeitsjahr von 90 Wochen zu je zwei Doppelstunden à 90 Minuten erreichen können.

Bei der Festlegung der Anforderungen soll eine klare Abgrenzung gegenüber den Prüfungen erfolgen, die Grundlage eines berufsständischen Nachweises für Dolmetscher und Übersetzer sind. Das *Higher Certificate* soll einen Kenntnisstand nachweisen, über den zum Beispiel junge Führungskräfte der Industrie verfügen müssen, um den Wirtschaftsteil einer Zeitung lesen, sich an Gesprächen mit ausländischen Geschäftsfreunden beteiligen und Routinegeschäftsbriefe schreiben zu können.

Besonderes Augenmerk soll den Bedürfnissen der Studenten der Volkswirtschaft und Technik und Hochschulen und Ingenieurschulen gewidmet werden.

Angesichts des Zeitdrucks und Zeitmangels, denen sich die oben zitierten Interessentenkreise vielfach ausgesetzt sehen, ist es unerläßlich, zur Intensivierung und Verkürzung des Lernprozesses den Einsatz von multimedialen Lehrwerken - programmiertes Übungsmaterial, audiovisuelle Hilfsmittel, Rundfunk und Fernsehen - von vornherein einzuplanen.

Neben der notwendigen Testentwicklung muß es Hauptanliegen der Grundlagenforschung sein, auf der Basis des bereits vorliegenden Mindestwortschatzes einen *Aufbauwortschatz Wirtschaft* zu erarbeiten. Die entsprechenden Forschungen und Analysen werden weit zeitraubender und umfangreicher sein als beim Mindestwortschatz, da auf diesem Gebiet nur geringe, unzureichende Forschungsergebnisse vorliegen.

Eine wesentliche Beschleunigung der Arbeiten könnte erzielt werden, wenn die Sichtung des umfangreichen Wortschatzes der Wirtschaft und Industrie durch zwei parallel tätige Arbeitsteams erfolgte, wobei die erste Gruppe untersuchen sollte, welche Wörter als zu fachspezifisch oder branchenbetont eleminiert werden können, während die zweite Gruppe jedes Item auf seine Unentbehrlichkeit untersuchen sollte. Im Interesse verläßlicher Arbeitsergebnisse wäre es erstrebenswert, von einem möglichst großen Corpus auszugehen und die Analyse durch Computerauswertung vorzunehmen.

Das Gesamtprojekt sollte unter Federführung des Bundesprüfungsbeauftragten des Deutschen Volkshochschul-Verbandes stehen und von den Mitgliedern des Internationalen Expertenausschusses getragen werden.

Das Projekt sieht eine Entwicklungsphase von 3 1/2 Jahren vor:

1. Jahr

Prüfungsordnung, genaue Definierung der Leistungsanforderungen der Prüfungsinhalte, Richtlinien für die Erarbeitung des Aufbauwortschatzes Wirtschaft, Darstellung der hierfür erforderlichen Forschungsmethoden.

2. Jahr

Erarbeitung des Aufbauwortschatzes Wirtschaft und erste Erprobungen, Testentwicklung im Fachbereich Wirtschaftsenglisch.

3. Jahr

Entwicklung von Diagnostik- und Einstufungstests sowie eines Modellsatzes von Prüfungsaufgaben zum Higher Certificate, Durchführung der ersten Prüfung als Modell.

3 1/2. Jahr

Itemanalyse der Prüfungsergebnisse und Erarbeitung eines Prüfungssatzes für die ersten bundeseinheitlichen bzw. internationalen Prüfungen.

Die Realisierung des oben dargestellten Projekts ist nunmehr ausschließlich eine Frage der Finanzierung. Die Mittel des Deutschen Volkshochschul-Verbandes reichen lediglich dazu, in den nächsten fünf Jahren die Entwicklung der Volkshochschul-Zertifikate *Russisch* und *Deutsch als Fremdsprache* durchzuführen und für die laufenden Prüfungstermine der bereits eingeführten Zertifikate *Englisch* und *Französisch* neue Prüfungssätze zu entwickeln.